Image and Vision Computing 145 (2024) 104997

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Siamese network to assess scanner-related contrast variability in MRI

Matteo Polsinelli [a,*], Hongwei Bran Li [b], Filippo Mignosi [c], Li Zhang [d], Giuseppe Placidi [e]

[a] Department of Management & Innovation Systems, University of Salerno, Italy
[b] Athinoula A. Martinos Center for Biomedical Imaging, Harvard Medical School, USA
[c] Department of Information Engineering, Computer Science, and Mathematics, University of L'Aquila, Italy
[d] People's Liberation Army Air Force Engineering University, China
[e] A²VI-Lab c/o Dept. MeSVA, University of L'Aquila, Italy

ARTICLE INFO

Keywords:
MRI
MRI pre-processing
Deep learning
Siamese network
Explainable AI

ABSTRACT

Magnetic Resonance Imaging (MRI) stands as a noninvasive tool for diagnosing and monitoring various diseases. The flexibility of MRI configuration parameters allows for adaptable imaging sequences, and at the same time poses challenges in terms of reproducibility, as variability in imaging sequences leads to significant differences in image contrast. This is one of the major causes that compromise the reliability of deep learning methods. Since the majority of the literature is focused on documenting the effects of this issue rather than delving into its underlying causes, this work follows a different approach. A Siamese Neural Network (SNN) has been trained to identify the scanner that acquired the input image. Experimental results include the use of Euclidean Distance (ED) and machine learning algorithms trained and tested using the feature vectors generated with the SNN. The results have shown that the proposed method is capable of distinguishing the scanner used for the acquisition with high accuracy. For a comprehensive interpretation of the results, the feature vectors have been dimensionality reduced and visualized with a 3D plot. Finally, the proposed method is sensitive to MR image contrast variability and could be used to detect data-related inconsistencies and provide a mechanism to make users aware of potential issues.

1. Introduction

Magnetic Resonance Imaging (MRI) is a noninvasive Medical Imaging (MI) tool for diagnosing and monitoring various diseases. The multitude of configuration parameters in MRI allows the production of adaptable imaging sequences, generating variable contrast images of different body tissues and organs and high-contrast images of the structures of interest [1,2].

However, this inherent flexibility poses challenges in terms of reproducibility [3,4] because variability in imaging sequences across scanners leads to significant differences in image contrast.

Furthermore, contrast variations can stem from dissimilarities among scanners, primarily attributed to magnetic field strength, acquisition protocols, manufacture, hardware imperfections, internal properties of the scanned body, and so on [2].

In Fig. 1 are shown several MR images (T1-Weighted) of the same subject available in the Dataset [5]. In particular, the subject has been scanned with 3 different scanners (named CMH, MRC, and ZHH) and each acquisition is repeated one year later. The image contrast differences between the scanners are clearly and also shown in the histograms calculated on the entire volume of the acquisition. Moreover, contrast differences are also visible between acquisitions from the same scanner CHM at years one and two.

Extensive training and testing on a comprehensive dataset covering various scanners are necessary to enhance generalization. However, acquiring a dataset containing all the possible contrast variations (or at the least the most representative ones) is unfeasible. Moreover, the manual annotation for each new acquisition is prohibitively expensive and difficult to implement. Additionally, this approach prolongs training time and complicates convergence.

In MI studies, especially the ones involving MRI, it is difficult to recruit participants (the patient population at any given center is limited) and the acquisitions are expensive and time-consuming. Consequentially, multi-site collaboration is a common practice to acquire the large samples needed and is fundamental for ensuring reliable and robust results [5]. However, in this scenario, the contrast variability becomes even more evident, presenting significant challenges for the effective use of the dataset. When supervised processing is used in MRI,

* Corresponding author.
E-mail address: mpolsinelli@unisa.it (M. Polsinelli).

https://doi.org/10.1016/j.imavis.2024.104997
Received 25 January 2024; Received in revised form 9 March 2024; Accepted 25 March 2024
Available online 29 March 2024

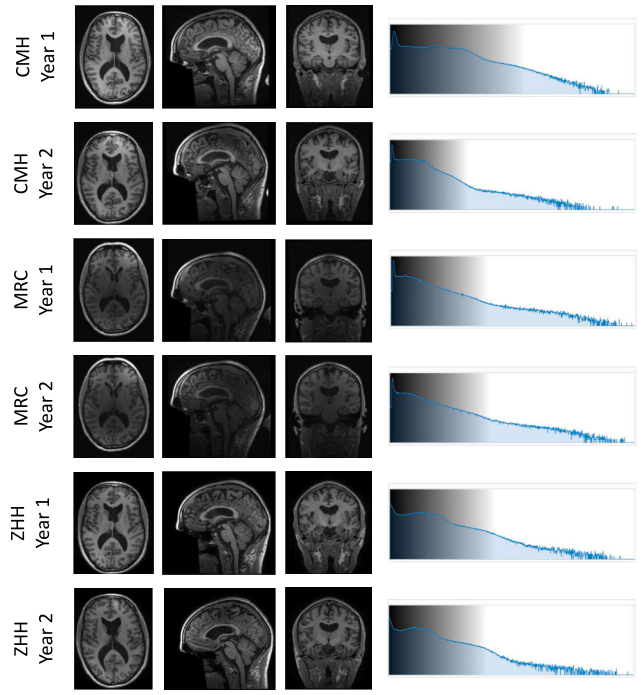

Fig. 1. Examples images taken from the Dataset [5]. The MRC, CHM, and ZHH are the labels assigned to each scanner in [5]. The images shown are referred to the subject 3. The histograms are calculated using the entire volume of each acquisition. Note: the histograms have different intensity scales, which have shown be extension of the gray zone.

huge and extensive labeled datasets are required to cover MRI variability, starting from contrast variation.

A more reasonable solution involves adopting pre processing strategies, specifically for contrast normalization. Existing literature extensively covers this direction and most of the proposed methods can be categorized into two distinct groups: statistical harmonization and DL harmonization [6].

For the first category, the ComBat method [7] is widely used to harmonize MRI and various extensions have been designed to optimize the method to different scenarios [6].

ComBat-style methods could in principle be applied for image-level harmonization at the voxel level, where images are co-registered and represented by vectorized voxel intensities, though not the recommended choice [6]. In this case, Deep Learning (DL) methods [8] could be a more reasonable choice. However, at the current state of the art, the end-user is warned to use them for several reasons: these methods have only been evaluated under ideal settings, may introduce unwanted and critical morphological tissue modification [9], and limited follow-up studies have been published for validating them with other datasets or in the clinical routine, leaving doubts about generalization [6]. Furthermore, to the best of our knowledge, an optimal algorithm for MRI harmonization does not exist [10].

The main consequence is that the DL models developed for specific MRI tasks, such as classification/segmentation of tissues/illness/lesions, trained on specific datasets may not perform optimally when applied to images from unseen scanners [11,12].

As shown in Fig. 1.a, usually the used Dataset is split between Trainin Dataset and Test Dataset. The Training Dataset is used by the DL model to learn the task until the convergence is reached. The Test Dataset is used to assess that the DL model is not overfitting. However, as shown in Fig. 1.b, this is not sufficient to ensure that the DL model will work well also when new images that differ from the Training Dataset are used as input.

Typically, studies tend to concentrate on documenting the effects of this issue rather than delving into its underlying causes. For example in [12], the impact of "MRI Manufacturer Shift" by employing three datasets comprising MR images from distinct scanners is evaluated. U-Nets trained for left ventricle segmentation exhibit high performance within their respective datasets but a noticeable drop in performance when applied to other datasets. A similar approach, utilizing adversarial learning, is presented in [13] for brain disorder classification.

Due to the limited literature on this subject [14–16], this work aims to propose an investigation of the impact of the contrast variation in MR images on the DL method's performance.

The aim of this work is threefold:

1. Demonstrating that contrast differences in MR images are such an evident phenomenon that can be used to train a neural network to recognize which scanner has generated the image. Since Siamese Neural Networks (SNNs) are trained to recognize differences in the images, they are a natural choice for this scope.
2. Demonstrating that even the trained SNN, exhibits a drop in performance for images for which the contrast variations are not represented in the Training Dataset.
3. Finally, it is discussed how the SNN could be included in an image processing pipeline to reduce the risks of image misclassification due to contrast variability.

This work is inspired by [17], where the aspect of MR image harmonization has been preliminary addressed. However, therein new experiments have been carried out, and a new point of view regarding the usage of the SNN in the image analysis pipeline is presented and discussed. Moreover, a notably greater number of techniques and methods to interpret the results have been used to significantly expand the experimentation.

New perspectives regarding the usage of SNN for MRI harmonization have been presented for future developments.

As in [17], the SNN [18] based on the EfficientNet Convolutional Neural Network (EN-CNN) [19], is trained to identify the scanner that acquired the input image. The SNN architecture is composed of three identical CNNs (same design and weights). During the training phase, the SNN learns the similarities between the anchor image and the positive image and learns the differences between the anchor image and the negative image. To this aim, the output of the SNN is a feature vector that is compared with the feature vectors of both the positive and negative examples using, for example, the Euclidean Distance (ED), as in Fig. 3.1.

For the first tests, the ED between the encoded vector of each test image and the encoded vector of all the training images has been calculated and used to evaluate the closest match to the scanner (as shown in Fig. 3.2).

To extend the experiments beyond the use of ED, the feature vectors output of the SNN are used as input to several machine learning algorithms [20–23], in particular, Support Vector Machines [24] (SVMs), Decision Trees [25] (DT), Random Forest [26] (RF), K-Nearest Neighbors [27] (KNN), and Logistic Regression [28] (LG), as shown in Fig. 3.3.

Finally, the proposed approach could be used to detect data-related inconsistencies, for example, detecting that the contrast of the input MR images is too different from the ones used for training and testing. For this reason, the proposed method could be used to make the final user aware of this inconvenience and take countermeasures to mitigate the problem. For example, fine-tuning the DL model before the usage (See Fig. 2).

For the experiments, the multi-scanner longitudinal multimodal MRI data set has been used [5], containing the acquisitions from several voluntary human subjects by multiple MRI scanners at different times.

2. Method

This section provides a comprehensive description of the

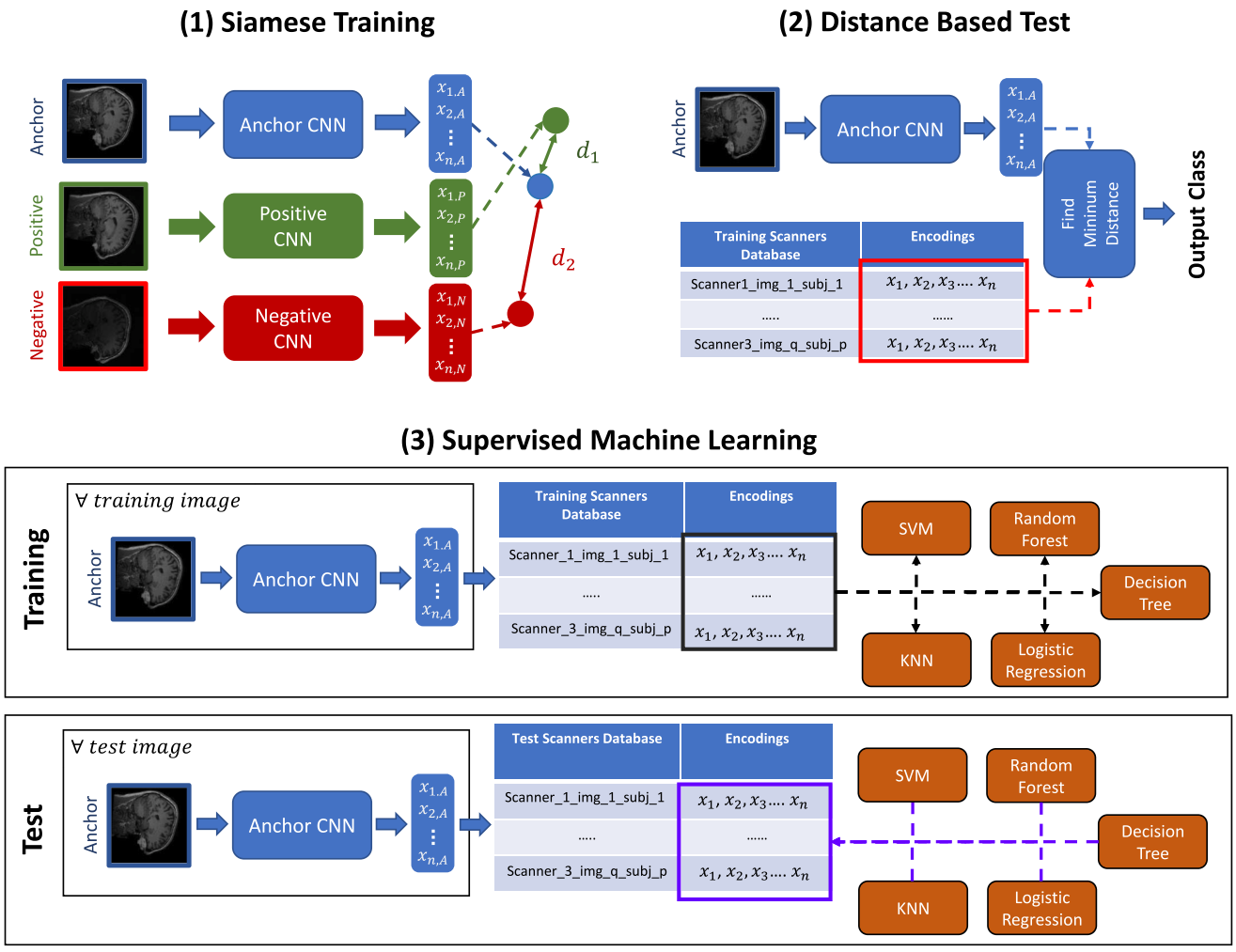

Fig. 3. The concept of the experiments carried out in this work.

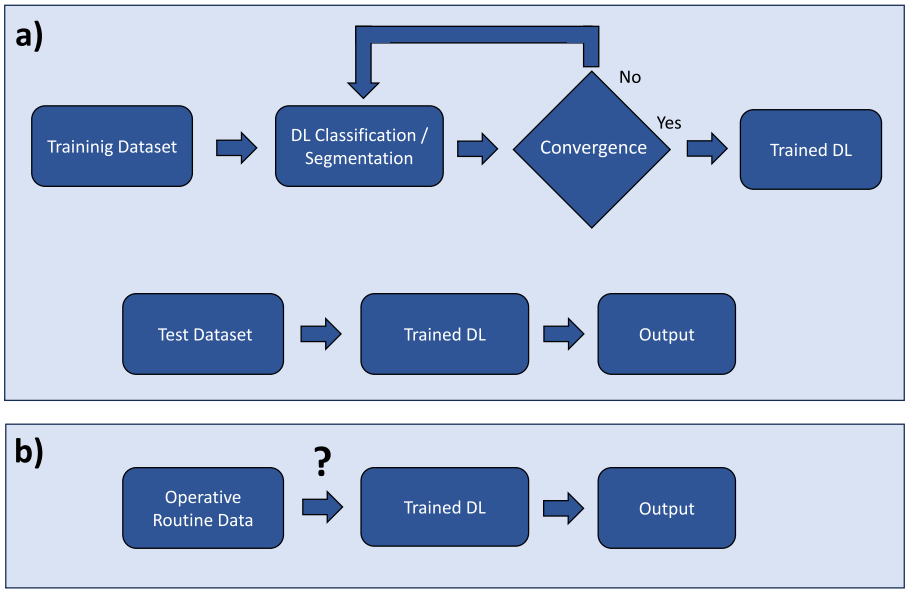

Fig. 2. The common processing pipeline used to Train and Test DL methods (a). Data inconsistency due to unseen MR images could lead to undesired effects on DL methods, like performance drop (b).

methodology and experiments. Moreover, it explains the rationale for the algorithms utilized in the work, justifying their selection within the context of the research objectives.

2.1. Siamese neural networks

SNNs are a class of neural network architectures that learn a similarity function, instead of learning to predict multiple classes. These architectures are called Siamese because they include two or more identical sub-nets (same configuration, parameters, and weights). SNNs are used to discover similarity by receiving three inputs, then extracting the features, and finally computing comparable feature vectors as output, in contrast with deep learning methods whose output is the belonging class.

In the context of MRI contrast variability ascertaining, the SNN is a reasonable choice because it can be trained to recognize which scanner has acquired the image. In this way, the SNN has to learn the particular intensity distribution that characterizes each scanner and encode these peculiarities to the output feature vectors. The generated feature vectors have the advantage that, in addition to being used for classification, they can be reduced and visualized to visually explain the behavior of the SNN, thus helping to understand how convolutional neural networks handle contrast variabilities.

During the training phase, the Triplet Loss [29] has been used. As the name suggests, the loss function compares a baseline (anchor) input with a positive (truthy) input and a negative (falsy) input. The main goal of an SNN is to minimize the distance of the baseline (anchor) input from the positive (truthy) input and to maximize the distance of the baseline input from the negative (falsy) input. This assumption aligns seamlessly with the purpose of this paper, for which the anchor is an MR image coming from a particular MRI scanner, the positive is an MR image from the same scanner and the negative is an MR image from a different scanner.

In the training process, an image triplet (Anchor, Positive, and Negative) is fed into the model as a single sample. With $image_{i,j}$ we refer to an image from the Training Dataset, where the first index is the scanner and the second index is the examined subject. The following Table 1, represents the conditions to which the training images undergo.

In summary, the resulting process must be scanner-dependent and subject-independent.

Defining the image triplet as (A, P, N), for each triplet q, the Triplet Loss is defined as:

$$L(A,P,N) = max\{d(A_q,P_q) + margin - d(A_q,N_q), 0\} \quad (1)$$

where the function d is defined as

$$d(x_q,y_q) = \| x_q - y_q \|_2 \quad (2)$$

and

$$margin > 0 \quad (3)$$

The *margin* parameter is used to 'stretch' the distance differences between similar and dissimilar pairs. In particular, three scenarios are handled with this loss:

- $d(A_i,N_i) > d(A_i,P_i) +$ *margin*: the negative image is significantly distant from the anchor image compared to the positive image. As the overall difference is negative, the loss function is 0, and there is no update to the SNN parameters.
- $d(A_i,N_i) < d(A_i,P_i)$: the negative image is in closer proximity to the anchor than the positive one. Consequently, the loss is positive, surpassing the specified *margin*, and leading to an update in the SNN parameters.
- $d(A_i,P_i) < d(A_i,N_i) < d(A_i,P_i) +$ *margin*: the negative image is farther from the anchor compared to the positive image, but the difference does not exceed the *margin*. Therefore, the loss remains positive (Even though smaller than the margin), resulting in an update to the SNN parameters.

Table 1
The input image triplet used for training the SNN.

$Anchor_{i,j}$	$\forall i, \forall j$
$Positive_{i,k}$	i and $\forall k$
$Negative_{s,k}$	$\forall s \neq$ i and $\forall k$

Each CNN used in the Siamese Architecture is an EfficientNet [19] model that is designed to achieve better accuracy and efficiency than previous ConvNets. EfficientNet achieves state-of-the-art accuracy on ImageNet Dataset, while being smaller and faster than the other CNNs [19].

The output feature vector dimension is 512. To test the SNN, one approach used in this work is based on the ED.

The ED, a metric quantifying the spatial separation between the feature vector of a test image and the feature vectors of all the training images, has been used to determine the closest match to a scanner and, hence, to determine the image classification.

2.2. Machine learning algorithms

For a more complete evaluation and investigation, different and more sophisticated approaches, beyond the ED-based, are required. Therefore, the output of the SNN has also been used to train and test several Supervised Machine Learning Algorithms to compare and discuss the classification results with the distance-based method.

In this work 5 different methods commonly used have been chosen: SVM, DT, RF, LG, and KNN.

SVM [30] is used for classification and regression problems. It aims to find a hyperplane in a high-dimensional space that best separates data points into different classes while maximizing the margin, which is the distance between the hyperplane and the nearest data points (support vectors). SVM does not work well when the dataset is noisy or contains several outliers.

Like SVM, DT is a supervised machine-learning approach used to solve both classification and regression problems. It recursively splits the dataset into subsets based on the most significant attribute at each node, creating a tree-like structure of decisions. It is sensitive to outliers, noise, and overfitting.

RF is the solution to overfitting in DT. This method is an ensemble of multiple DT and, besides overfitting, it is also more robust to outliers and noise than DT. Moreover, RF is more suitable for multi-dimensional data.

A simpler method that can be used for both classification and regression, is KNN. For classification, the algorithm counts the class labels of the k neighbors and assigns the class label with the highest count to the new data point. However, k-NN doesn't perform well on imbalanced data and is very sensitive to outliers as it simply selects the neighbors relying on distance-based criteria.

The last method we consider is the LG. The algorithm is mainly used for binary classification problems. It is relatively robust to noise but it cannot used to solve non-linear problems because it has a linear decision surface.

To simplify the discussion, Table 2 reports the fundamental aspects of the considered methods.

2.3. Evaluation metrics

Evaluation metrics provide a comprehensive assessment of the performance of a classification model while allowing a comparison between different models. In this work, Precision, Recall, Specificity, and F1 scores are calculated and used for comparison. Since the role of these metrics is fundamental to understanding the behavior of the SNN, it is

Table 2
Comparison between machine learnings methods used in this work.

Method	Type of Algorithm	Underlying Concept	Training Approach	Sensitivity
SVM	Supervised learning algorithm for classification and regression	Finds the hyperplane that best separates different classes in a high-dimensional space	Finds the optimal hyperplane through the training data	Normal to Outliers and noise
DT	Both classification and regression	Tree-like structure where each node represents a decision based on a feature	Builds a tree structure based on feature splits	Potential to outliers
RF	Ensemble learning method for classification and regression	Ensemble of decision trees, combines their predictions for improved accuracy	Trains multiple decision trees on random subsets of the data	Low to outliers
KNN	Both classification and regression	Classifies a data point based on the majority class of its k-nearest neighbors	Classifies based on proximity	High to outliers and to unbalanced data
LG	Classification	Models the probability of a binary outcome using a logistic function	Optimizes parameters using the logistic function	Normal to outliers low to noise

necessary to define them. Assuming that True Positives are reported as TP, True Negatives are reported as TN, False Positives are reported as FP, and False Negatives are reported as FN, the formulas for these metrics are the following:

$$Prec = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Spec = \frac{TN}{TN + FP} \tag{6}$$

$$F1 = \frac{2^{*} Prec^{*} Recall}{Prec + Recall} \tag{7}$$

In particular, Precision (Prec) concerns the prediction accuracy of the TP, focusing on minimizing FP, and Recall also concerns the accuracy of the TP but minimizing FN.

The increment of Prec often leads to a decrease in Recall and vice versa. An optimal balance between Prec and Recall depends on the specific requirements of the application. For example in medical diagnosis, high Recall might be more important than high Prec.

Since Prec and Recall focus on the positive class, Spec. is also reported because it focuses on the correct identification of the TN.

F1 provides a single metric considering both Prec and Recall.

2.4. Contrast normalization

As previously stated, this work aims to demonstrate that contrast variation is indeed the primary factor leading a DL method astray. To confirm that, it is necessary to standardize the images, repeat the tests, and observe the results. If improvements are observed, it would provide further grounds to attribute the variations to the variable contrast. Again, there are a large number of techniques that could be used [10,31–33], though no one optimal method exists. We use Histogram Matching [34] (HM) to align the, approximately linear, contrast variation in MR images from the same scanner, occurring at different times. Due to the limited and linear contrast variation, HM is a good solution for image harmonization (this is not true when non-linear contrast variations occur, as in the case of different scanner comparisons). The goal here is to make their histograms more similar, resulting also in good a visual resemblance between the images.

3. Experimental design and setup

3.1. The used data set

For this study, the longitudinal multimodal MR image dataset [5] has been used. The dataset is designed to assess contrast variation between scanners, and the robustness of the pipeline results, and to examine approaches for harmonization across multiple scanners, while also accounting for changes in body anatomy over time.

The data set consists of acquisitions from three different MRI scanners: a General Electric 750w Discovery 3 T labeled as CMH, a Siemens Tim Trio 3 T labeled as MRC, and a General Electric 750 Signa 3 T labeled as ZHH. Moreover, are available also the acquisition of three Siemens Prisma 3 T MRI scanners, located in three different centers and labeled as CMP, MRP, and ZHP.

Four voluntary subjects were selected for data collection and acquisitions were made at different times in 4 years.

3.2. Data setup

To study the effects of MR images from different MRI scanners, the SNN has been trained on the acquisitions of the first year, subjects 1, 3, and 4 (subject 2 is only available from the third year), and tested on the same subjects and scanners in year 2, using the T1-weighted (T1w) modality.

In this way, the Training Dataset was composed of 1539 MR images the same number as the Test Dataset. Each image is a greyscale image of 256*256 pixels.

As observed in [17], the histograms of each scanner in year 1, grouped by subjects, show relevant differences in terms of amplitude distribution. Moreover, regarding the histograms of each scanner in year 2, still grouped by subjects, it can be observed that while the scanners MRC and ZHH maintain almost unaltered their original distribution of year 1, CMH slightly deviates from its original shape.

For this reason, we expected that during the test phase, the scanner that is the most difficult to recognize is the CMH since its behavior changed, while the MRC and ZHH mostly remained unaltered.

It is important to remark that the dataset contains some repeated acquisitions for the same subject with the same scanner. Looking at their histogram, no significant variations need to be reported and regarding the acquisition with the CHM scanner, the deviations previously discussed are still present. Differences among these repetitions could be related to motion artifacts that are visible in some images. However, this does not represent an issue for the scope of our work.

3.3. Experimental setup

All experiments have been executed on a personal computer with the following hardware configuration: AMD Ryzen 76800H, Nvidia RTX 3060, 16 GB RAM, 1 TB SSD, and Windows 11 OS. The software environment used was Python 3.10 and the Jupyter Notebook. PyTorch [35] has been used to implement the SNN and Scikit-Learn [36] to use the dimensionality reduction methods, to train and test the machine learning algorithms, and to compute the confusion matrices and performance metrics.

4. Results and discussion

4.1. Training

To train the SNN, the Adam optimizer has been used with the following hyperparameters: batch size 1, learning rate 0.0005, and training epochs 25. The SNN was trained from scratch without finetuning.

The hyper-parameters of the machine learning methods used in this work are the following:

- for the K-NN the number of neighbors is equal to 3;
- for the SVM the kernel type is radial basis function, and the regularization parameter is equal to 1.0;
- for the DT the criterion is Gini impurity, and the splitting rule is "best";
- for the RF the number of trees is 100, the max depth is 2 and the criterion is Gini impurity;
- for the LR the penalty is "l2", the maximum number of iterations is 100 and the solver is the Limited-memory Broyden–-Fletcher–Goldfarb–Shanno Algorithm.

The SNN is capable of classifying correctly all the training images, without making any misclassification at all. Likewise, all the machine learning algorithms trained on the training encoded feature vectors showed the same capabilities.

4.2. Testing

The averaged results for all the classes of the test images are reported in Table 3.

The first row reports the configuration SNN and Euclidean Distance while the remaining rows report the configurations SNN and the machine learning algorithms.

On average, the best combination is represented by SNN + SVM which outperforms all the remaining combinations for all the metrics.

SNN + LG achieves the second-best classification performance, even if it is better than SNN + SVM to capture the non-linearity in the data and more robust to outliers. A possible explanation is that the scanner's classes can be well separated by hyperplanes.

Interestingly, the commonly used combination SNN + ED is, in general, never the best.

Notably, as we expected, on average SNN + RF outperforms SNN + DT, being RF an ensemble of DT and is more robust to outliers.

Another consideration that emerges from the results in Table 3 is that for all the investigated configurations the Specificity is the score with the higher value while both Prec and Recall exhibit lower values.

For this reason, a deep investigation is necessary and the results scores obtained maintaining the 3 classes separately are shown in Table 4.

Regarding Prec, all the methods were capable of classifying CMH better than MRC and ZHH. In both MRC and ZHH classes, the number of FP is higher because their clusters contain elements from the CMH cluster.

Regarding Recall, almost all the methods showed very good performance for MRC and ZHH compared to CMH. The CMH cluster has several FNs, baing most of its elements absorbed by MRC and ZHH.

Regarding Spec, the results are a bit more fair. In general, CMH achieves better results but the differences with the other classes are not so marked. Has previously stated, MRC and ZHH clusters have more FP compared to CMH but less TN.

Notably, the combination SNN + ED achieves a very good F1-Score for the class CMH but, at the same time, the F1-score for the MRC is the worst one and the F1-score for the ZHH is the second worst one. This suggests that also this combination is not capable of well separating the images.

Table 3
The averaged results for all the test images. The values are in %.

Method	Pre	Recall	Spec	F1
SNN + ED	88.30	90.61	95.02	87.71
SNN + SVM	**94.12**	**93.31**	**96.65**	**93.15**
SNN + DT	85.40	84.73	92.37	84.49
SNN + RF	91.17	90.38	95.19	90.10
SNN + KNN	90.54	88.11	94.05	87.50
SNN + LG	92.46	91.36	95.68	91.01

4.3. Testing with contrast normalized

As previously discussed, the MR images acquired from the CMH scanner in year 2 (that are part of the Test Dataset) have a different contrast compared to year 1 (that are part of the Training Dataset). Consequentially the SNN exhibits a drop in classification performance during the test phase. For this reason, using histogram matching to make the histogram of the images of year 2 similar to the images of year 1 (given a subject's year 2 image, is it matched to the same subject's year 1) should improve the classification performance.

The averaged results for all the test images are reported in Table 5. Remarkably, all the method have significantly improved their performance, meaning that histogram matching has been considerably effective. On average, the best combination is still represented by SNN + SVM which outperforms all the remaining combinations for all four metrics.

For the sake of completeness, Table 6 reports the results scores obtained maintaining the 3 scanners separately. The behavior of the SNN in combination with other methods is the same as discussed in Section 4.2.

From the obtained results, it would seem that histogram matching is easily solving the harmonization problem. It is advisable to bear in mind that this has proven to be true for this particular case of study, but in general, the task is not as straightforward at all.

In fact, in this case study the contrast variabilities seem to be uniformly distributed across the entire image and for this reason, histogram matching is effective. Rather, it manifests disparately across distinct anatomical regions [32], leading to differential impacts on image quality and clarity in different areas.

Moreover, for real complex tasks, like image classification for disease diagnosis or tissue segmentation, one can expect that the effects of contrast variability can be even more evident, and and the same time, contrast harmonization will be even more difficult, and more advanced techniques will be necessary and fundamental for the successful outcome.

4.4. Visual evaluation

For a comprehensive analysis of the results, it is essential to visualize the feature vectors generated by the SNN. Therefore, it is necessary to reduce the dimensions of the encoded feature vector from 512 to a 3D Cartesian space, where a 3D plot can be visualized. This requires the use of algorithms for feature dimension reduction.

The goal of these kinds of methods is usually to reduce the number of input variables while trying to retain the relevant information. Several approaches are available [37] but in this work, we use three different feature reduction methods: Principal Component Analysis [38] (PCA), Independent Component Analysis [39] (ICA), and Uniform Manifold Approximation and Projection [40] (UMAP).

PCA works by finding the principal components of the data, which are the directions of maximum variance. The data is projected onto a lower-dimensional space that captures as much of the variance as possible while preserving the overall structure. However, PCA struggles with complex nonlinear relationships between variables and assumes that the data is Gaussian-distributed.

Table 4
The classes MRC, CMH, and ZHH represent the labels of the scanners reported in [5]. The values are in %.

Method	Prec			Recall			Spec			F1		
	MRC	CMH	ZHH	MRC	CMH	ZHH	MRC	CMH	ZHH	MRC	CMH	ZHH
SNN + ED	64.91	100.00	100.00	100.00	94.47	77.37	85.07	100.00	100.00	78.72	97.16	87.24
SNN + SVM	96.43	100.00	85.93	100.00	79.92	100.00	98.15	100.00	91.81	98.18	88.84	92.43
SNN + DT	92.71	87.16	76.35	94.15	68.81	91.23	96.30	94.93	85.87	93.42	76.91	83.13
SNN + RF	94.07	96.22	83.22	99.03	74.46	97.66	96.88	98.54	90.16	96.49	83.96	89.87
SNN + KNN	94.82	100.00	76.80	100.00	64.33	100.00	97.27	100.00	84.89	97.34	78.29	86.88
SNN + LG	92.60	100.00	84.79	100.00	74.07	100.00	96.00	100.00	91.03	96.16	85.11	91.77

Table 5
The averaged results for all the test images. The values are in %.

Method	Pre	Recall	Spec	F1
SNN + ED	99.48	99.48	99.74	99.47
SNN + SVM	**100.00**	**100.00**	**100.00**	**100.00**
SNN + DT	88.20	88.11	94.05	88.04
SNN + RF	93.33	92.98	96.49	92.87
SNN + KNN	99.42	99.42	99.71	99.41
SNN + LG	99.29	99.29	99.64	99.28

Like PCA, ICA is a linear method used in data analysis to separate a multivariate signal into additive, independent components. However, while PCA focuses on finding uncorrelated components, the objective of ICA is to find a linear transformation of the observed data such that the resulting components are as statistically independent from each other as possible.

UMAP [40] is a dimensionality reduction technique known for its effectiveness in preserving both local and global structures in the data. Compared with PCA and ICA, UMAP is a non linear method and for this reason, it is particularly useful for capturing complex relationships in high-dimensional data.

The presented methods are reported in Table 7 for comparison.

These methods have been used for the feature vectors of the training images, test images, and test images corrected through histogram matching. The plots are reported in Fig. 4.

Regarding the reduced feature vectors generated from the training images, the plots are reported in Figs. 4.a, 4.b, and 4.c. In general, for the encoded features vector UMAP, there is no overlapping and the 3 groups of scanners are well separated (Fig. 4.c). Regarding PCA and ICA, the 3 groups are still well separated except for some uncommon outliers which can be attributed to the strong reduction of features (Figs. 4.a and 4.b).

Regarding the reduced feature vectors generated from the test images, several outliers are present in the features reduced with PCA, ICA, and UMAP (Fig. 4.d, 4.e, 4.f).

In particular, the outliers are represented by several points that have shifted from the CMH cluster towards both the MRC and ZHH clusters. In other words, several images acquired from CMH, for the SNN, are similar to the images from MRC and ZHH.

Finally, the reduced feature vectors generated from the test images after histogram matching are shown in Figs. 4.g,4.h, 4.i. and some outliers are still present. Compared to Figs. 4.d, 4.e, and 4.f, their number has been significantly reduced but not enough to justify the performance improvement reported in Table 4. However, it is necessary to keep in mind that the reduction methods are reducing the feature vectors from 512 elements to just 3 elements and it is permissible to assume that outliers should be attributable to this process. Moreover, it is out of the scope of this work to quantify the numbers of outliers before the harmonization and after the harmonization using reduction methods. What was expected and was obtained is improved clustering with a notably reduced number of outliers.

5. Updated pipeline

The obtained results have demonstrated that the SNN is capable of generating feature vectors that well separate sets of MR images acquired with different scanners and, consequentially, recognize which scanner has generated the input image with high precision.

However, suppose that a new subject is scanned with an MRI scanner that, for the reasons previously discussed, produces images with a contrast that is not represented in the Training Dataset. In this case, the SNN will be not capable of associating each MR image with the same scanner label, as discussed in Section 4.2. The results will be that the SNN will classify the images with different labels, and this behavior could be used to detect the MRI acquisition with different contrasts.

Table 6
The classes MRC, CMH, and ZHH represent the labels of the scanners reported in [5]. All the values have to be intended in %.

Method	Prec			Recall			Spec			F1		
	MRC	CMH	ZHH	MRC	CMH	ZHH	MRC	CMH	ZHH	MRC	CMH	ZHH
SNN + ED	98.44	100.00	100.00	100	99.42	99.03	99.23	100	100	99.21	99.71	99.51
SNN + SVM	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
SNN + DT	92.75	88.34	83.51	94.74	79.73	89.86	96.30	94.74	91.13	93.73	83.81	86.57
SNN + RF	95.85	96.57	87.59	99.03	82.26	97.66	97.86	98.54	93.08	97.41	88.84	92.35
SNN + KNN	98.84	100.00	99.42	100.00	98.25	100.00	99.42	100.00	99.71	99.42	99.12	99.71
SNN + LG	98.65	100.00	99.23	100.00	97.86	100.00	99.32	100.00	99.61	99.32	98.92	99.61

Table 7
Comparison between feature reduction methods used in this work. Information collected by [41].

Method	Objective	Linearity	Orthogonality	Weakness	Outliers	Applications
PCA	Maximize Variance	Linear	Yes	Limited to linear projection	Sensitive	Dimensionality reduction Feature extraction
ICA	Maximize statistical independence	Linear	Not Required	Require high training time	Sensitive	Blind source separation
UMAP	Preserves both local and global structures	Non-linear	No	Sensitivity to hyperparameter choices	Sensitive	Data visualization Classification

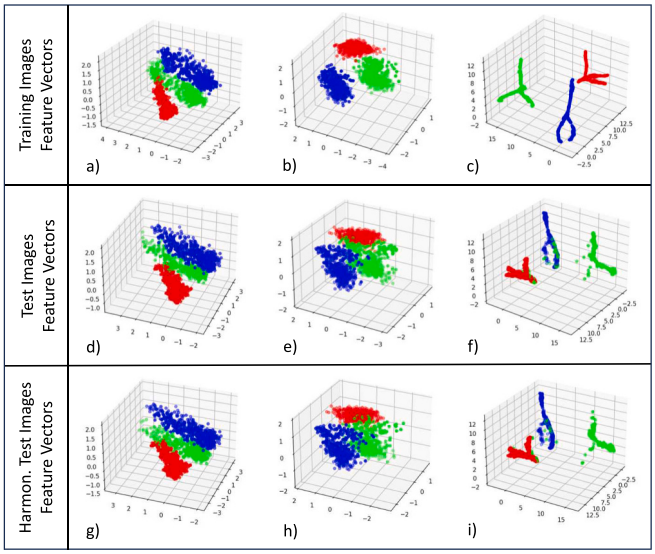

Fig. 4. SNN encoded features of training images: a) PCA, b) ICA, c) UMAP. SNN encoded features of test images: d) PCA, e) ICA, f) UMAP. SNN encoded features of harmonized test images: g) PCA, h) ICA, i) UMAP. The red, green, and blue represent respectively the MRC, CHM, and ZHH scanners according to the label assigned in [5]. Each plotted point is related to an MR image given as input to the SNN that generates the feature vector of 512 elements and then reduced to the 3 dimensions using the discussed features reduction algorithms. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

This behavior could be used for detecting MR images that exhibit significant contrast variations compared to the training data. Fig. 5 is shown as a possible new processing pipeline that improves the common one shown in Fig. 1.

In the new pipeline introduce a new step that checks the MR images, using the SNN, before these last ones are fed to the DL models.

Suppose that there are n different labels of scanners inside the Training Dataset. Each image of a new MRI acquisition is classified with the SNN, and iteratively, their label is chosen between all the n labels available. If the SNN is capable of associating almost all the new MR images to the same scanner label, it is possible to conclude that the contrast of the images is similar to the ones presented in the Training Dataset. In this case, no further actions are required. On the contrary, if the images are classified into several different labels, it means that the SNN has not seen the contrast of the image because cannot find a representative scanner (in the Training Dataset) for which the contrasts of the images are similar. Consequentially, further actions are recommended, like proceeding with Fine-Tuning both the SNN and the DL model as shown in Fig. 5.

6. Conclusions and future work

Acquiring an in-depth comprehension of how a CNN manages these variations could serve as an initial step to enable DL methods to be robust and maintain performance even on images acquired with unseen scanners. Moreover, it could serve to develop more efficient image harmonization techniques and could provide valuable advice for the data collection phase.

In this work, we present an investigation of the effects of MR image variabilities during a CNN's training and test phase, in this case, an SNN.

The SNN was trained on MRI images acquired from three different subjects and three different MRI scanners to recognize which scanner generated the input MR image. Since the output feature vectors of the SNN cannot be used directly for classification, the Euclidean distance has been used to calculate the distance between the feature vector of the input image with all the feature vectors of the training images to find the correct class. To go beyond the constraints of this technique, five different machine learning algorithms have been trained on the output feature vectors of the SNN to be used in combination with the SNN. The results show that the network is learning to recognize them.

The test phase has also confirmed that the SNN is capable of recognizing MR images, especially from the scanners MRC and ZHH. Interestingly, the MR images from the scanner CMH are the ones more affected by misclassification and it seems related to the histogram shift that occurred between year 1 and year 2.

Finally, using histogram matching, the CMH MR images have been harmonized concerning the same scanner of year 1 (used for training). The classification results have significantly improved as was expected.

It is important to notice that histogram matching has proven to be effective for this particular case of study. However, real-scenario tasks are more complex, and one can expect that the effects of contrast variability can be even more evident, and at the same time, contrast harmonization will be even more difficult [32].

To visually evaluate the output feature vectors generated by the SNN, they have been projected into a 3D space using PCA, ICA, and UMAP. The plots relative to the training images have shown three distinct clusters, one for each MRI scanner, which confirms that there are considerable differences in how the SNN encoded those features. Instead, regarding the plots that report the reduced feature vectors of the test images, several elements of the cluster of the CMH scanner have moved to the clusters MRC and ZHH, further affirming the numerical results. Finally, observing the plots of the reduced features of the test images after histogram matching is applied, it is possible to notice that the outlier elements relative to the CHM scanner are drastically reduced.

Finally, it has been discussed how the SNN could be used as a method for detecting contrast variability in MR images, increasing the robustness of the processing pipelines in general and in particular the ones that use DL methods.

With this aim, for future development, this scenario will be carried on. First of all, the SNN will be trained to detect contrast variabilities also for different MR modalities, like Proton Density, T2-weighted, and

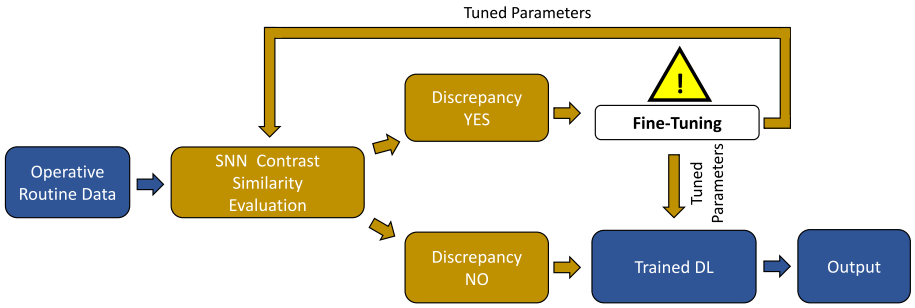

Fig. 5. The proposed solution to mitigate the contrast variability effects on DL models. Compared with Fig. 1, a new step is introduced between the input data and the Trained DL. This step, using the SNN, is used to detect discrepancies between MR images used during training and testing and the new MR images.

Fluid-attenuated inversion recovery (FLAIR), beyond the T1 used in this work. Subsequently, the discussed pipeline will be tested with several multimodal, multicenter, multi-scanner datasets used to train DL methods for tasks like multiple sclerosis lesion segmentation, tumor detection, etc. to assess the behavior and the performances of the SNN also with MR images acquired in the presence of abnormal tissues such as brain tumor.

CRediT authorship contribution statement

Matteo Polsinelli: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Conceptualization. **Hongwei Bran Li:** Writing – review & editing. **Filippo Mignosi:** Writing – review & editing. **Li Zhang:** Writing – review & editing. **Giuseppe Placidi:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] M.A. Bernstein, K.F. King, X.J. Zhou, Handbook of MRI Pulse Sequences, Elsevier, 2004.
[2] G. Placidi, MRI: Essentials for Innovative Technologies, CRC Press, 2012.
[3] B.E. Dewey, C. Zhao, J.C. Reinhold, A. Carass, K.C. Fitzgerald, E.S. Sotirchos, S. Saidha, J. Oh, D.L. Pham, P.A. Calabresi, et al., Deepharmony: a deep learning approach to contrast harmonization across scanner changes, Magn. Reson. Imaging 64 (2019) 160–170.
[4] G. Placidi, L. Cinque, M. Polsinelli, Guidelines for Effective Automatic Multiple Sclerosis Lesion Segmentation by Magnetic Resonance Imaging, ICPRAM, 2020, pp. 570–577.
[5] C. Hawco, E.W. Dickie, G. Herman, J.A. Turner, M. Argyelan, A.K. Malhotra, R. W. Buchanan, A.N. Voineskos, A longitudinal multi-scanner multimodal human neuroimaging dataset, Sci. Data 9 (1) (2022) 332.
[6] F. Hu, A.A. Chen, H. Horng, V. Bashyam, C. Davatzikos, A. Alexander-Bloch, M. Li, H. Shou, T.D. Satterthwaite, M. Yu, et al., Image harmonization: a review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization, NeuroImage 120125 (2023).
[7] J.-P. Fortin, D. Parker, B. Tunç, T. Watanabe, M.A. Elliott, K. Ruparel, D.R. Roalf, T. D. Satterthwaite, R.C. Gur, R.E. Gur, et al., Harmonization of multi-site diffusion tensor imaging data, Neuroimage 161 (2017) 149–170.
[8] L. Di Biasi, F. De Marco, A. Auriemma Citarella, M. Castrillón-Santana, P. Barra, G. Tortora, Refactoring and performance analysis of the main cnn architectures: using false negative rate minimization to solve the clinical images melanoma detection problem, BMC Bioinform. 24 (1) (2023) 386.
[9] J.P. Cohen, M. Luck, S. Honari, Distribution matching losses can hallucinate features in medical image translation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I, Springer, 2018, pp. 529–536.
[10] L.J. Isaksson, S. Raimondi, F. Botta, M. Pepa, S.G. Gugliandolo, S.P. De Angelis, G. Marvaso, G. Petralia, O. De Cobelli, S. Gandini, et al., Effects of mri image normalization techniques in prostate cancer radiomics, Phys. Med. 71 (2020) 7–13.
[11] G. Placidi, L. Cinque, M. Polsinelli, A. Splendiani, E. Tommasino, Automatic framework for multiple sclerosis follow-up by magnetic resonance imaging for reducing contrast agents, in: Image Analysis and Processing–ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part II 20, Springer, 2019, pp. 367–378.
[12] W. Yan, L. Huang, L. Xia, S. Gu, F. Yan, Y. Wang, Q. Tao, Mri manufacturer shift and adaptation: increasing the generalizability of deep learning segmentation for mr images acquired with different scanners, Radiol.: Artif Intell. 2 (4) (2020) e190195.
[13] H. Guan, Y. Liu, E. Yang, P.-T. Yap, D. Shen, M. Liu, Multi-site mri harmonization via attention-guided deep domain adaptation for brain disorder identification, Med. Image Anal. 71 (2021) 102076.
[14] B. Glocker, R. Robinson, D. C. Castro, Q. Dou, E. Konukoglu, Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects, arXiv preprint arXiv:1910.04597.
[15] W.M. Kouw, M. Loog, L.W. Bartels, A.M. Mendrik, Learning an mr acquisition-invariant representation using siamese neural networks, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019, pp. 364–367.
[16] H. Mi, M. Yuan, S. Suo, J. Cheng, S. Li, S. Duan, Q. Lu, Impact of different scanners and acquisition parameters on robustness of mr radiomics features based on women's cervix, Sci. Rep. 10 (1) (2020) 20407.
[17] M. Polsinelli, L. Cinque, F. Mignosi, G. Placidi, G. Tortora, Siamese network to investigate scanner-dependency in mri, in: 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2023, pp. 535–541.
[18] L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, P.H. Torr, Fully-convolutional siamese networks for object tracking, in: Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part II 14, Springer, 2016, pp. 850–865.
[19] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.
[20] I.H. Sarker, Machine learning: algorithms, real-world applications and research directions, SN Comp. Sci. 2 (3) (2021) 160.
[21] A. Singh, N. Thakur, A. Sharma, A review of supervised machine learning algorithms, in: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), Ieee, 2016, pp. 1310–1315.
[22] F. De Marco, F. Ferrucci, M. Risi, G. Tortora, Classification of qrs complexes to detect premature ventricular contraction using machine learning techniques, PLoS One 17 (8) (2022) e0268555.
[23] F. De Marco, L. Di Biasi, A.A. Citarella, M. Tucci, G. Tortora, Identification of morphological patterns for the detection of premature ventricular contractions, in: 2022 26th International Conference Information Visualisation (IV), IEEE, 2022, pp. 393–398.
[24] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, IEEE Intel. Syst. Appl. 13 (4) (1998) 18–28.
[25] W.-Y. Loh, Classification and regression trees, Wiley Int. Rev. Data Min. Knowledge Disc. 1 (1) (2011) 14–23.
[26] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[27] A. Mucherino, P.J. Papajorgji, P.M. Pardalos, A. Mucherino, P.J. Papajorgji, P. M. Pardalos, K-nearest neighbor classification, Data Min. Agricult. (2009) 83–106.
[28] D.W. Hosmer Jr., S. Lemeshow, R.X. Sturdivant, Applied Logistic Regression Vol. 398, John Wiley & Sons, 2013.
[29] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
[30] L. Wang, Support Vector Machines: Theory and Applications vol. 177, Springer Science & Business Media, 2005.
[31] R.T. Shinohara, E.M. Sweeney, J. Goldsmith, N. Shiee, F.J. Mateen, P.A. Calabresi, S. Jarso, D.L. Pham, D.S. Reich, C.M. Crainiceanu, et al., Statistical normalization techniques for magnetic resonance imaging, NeuroImage: Clin. 6 (2014) 9–19.
[32] G. Placidi, M. Polsinelli, Local contrast normalization to improve preprocessing in mri of the brain, in: Bioengineering and Biomedical Signal and Image Processing: First International Conference, BIOMESIP 2021, Meloneras, Gran Canaria, Spain, July 19–21, 2021, Proceedings 1, Springer, 2021, pp. 255–266.
[33] M. Liu, P. Maiti, S. Thomopoulos, A. Zhu, Y. Chai, H. Kim, N. Jahanshad, Style transfer using generative adversarial networks for multi-site mri harmonization, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24, Springer, 2021, pp. 313–322.
[34] R. C. Gonzalez, R. E. Woods, Digital image processing. upper saddle river, J.: Prentice Hall.
[35] PyTorch, https://pytorch.org/, [Online; accessed 16/01/2024] (2024).
[36] Scikit-Learn, https://scikit-learn.org/stable/, [Online; accessed 16/01/2024] (2024).
[37] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, J. Saeed, A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction, J. Appl. Sci. Technol. Trends 1 (2) (2020) 56–70.
[38] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1–3) (1987) 37–52.
[39] G. R. Naik, D. K. Kumar, An overview of independent component analysis and its applications, Informatica 35 (1).
[40] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426.
[41] F. Anowar, S. Sadaoui, B. Selim, Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, Ica, t-sne), Comput. Sci. Rev. 40 (2021) 100378.