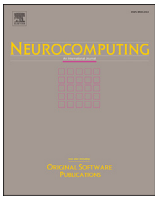
Neurocomputing 692 (2026) 133854

Contents lists available at ScienceDirect

Neurocomputing

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



ReSHAPe: A redundancy-reduced SHAP-based feature selection pipeline for interpretable radiomics in biomedical image analysis

Alessandra Perniciano*, Federico Maria Cau, Lucio Davide Spano, Cecilia Di Ruberto, Andrea Loddo*

Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, 09124, Cagliari, Italy

ARTICLE INFO

Communicated by D. Liu

Keywords:
Biomedical imaging
Radiomics
Interpretable machine learning
Feature selection
SHAP

ABSTRACT

Handcrafted radiomic descriptors are widely used in biomedical image analysis, but classic radiomics pipelines often suffer from high feature redundancy and an underdeveloped, weakly principled feature-selection practice, which together can impair generalization and limit model interpretability. To address this, we introduce ReSHAPe (Redundancy-Reduced SHAP-based Evaluation), a two-stage, model-aware feature selection pipeline that makes SHAP-driven selection practical for radiomics. ReSHAPe first performs redundancy pruning by removing highly correlated features using Spearman rank correlation, retaining within each correlated group the descriptor with the lower absolute skewness. It then applies SHAP-based global importance to rank the remaining features and iteratively select a compact subset; an ensemble variant aggregates SHAP rankings across multiple classifiers to promote consensus and interoperability.

We evaluate ReSHAPe on three MedMNIST v2 subsets (BreastMNIST, PneumoniaMNIST, BloodMNIST) using 285 handcrafted features and five well-known classifiers (SVM, Decision Tree, Random Forest, Extra Trees, XGBoost), comparing against univariate filters (ANOVA F-test, mutual information), SHAP-only selection, correlation-based filtering, and full-feature baselines. Across datasets, ReSHAPe preserves performance while drastically reducing dimensionality; on radiomic tasks, it is consistently competitive with SHAP-only selection f-measure weighted values differences typically lower than 0.03, and it remains effective in the non-radiomic multiclass setting (maximum decrease of f-measure weighted value lower than 0.04). Finally, the correlation pre-filtering stage markedly reduces SHAP overhead, which would otherwise require 200 additional model training/evaluation steps when applied directly to the full feature space.

1. Introduction

The analysis of biomedical images is of fundamental importance in contemporary healthcare, as it supports diagnosis, prognosis, and treatment planning. Machine learning (ML) and deep learning (DL) methods have shown strong potential in the interpretation of X-rays, magnetic resonance imaging (MRI), and pathology slides. However, widely adopted benchmarks such as MNIST and ImageNet were not designed to capture the heterogeneity and complexity of biomedical data [1]. In response, several standardized medical imaging datasets have been introduced to provide accessible benchmarks with consistent preprocessing, rich annotations, and diverse modalities [2].

Within this landscape, radiomics has emerged as a powerful paradigm for quantitatively characterizing medical images by extracting large sets of handcrafted features that describe intensity, shape, and texture properties of regions of interest [3–5]. These descriptors are intended to encode phenotypic manifestations of underlying pathophysiological processes and have been applied to tasks such as lesion detection, grading, outcome prediction, and response assessment. Yet radiomic feature spaces are typically high-dimensional and highly redundant [6], which can impair generalization, increase computational costs, and limit the interpretability of downstream predictive models.

The MedMNIST family of datasets [1,2] is a lightweight yet diverse benchmark suite for biomedical image analysis, spanning multiple modalities and anatomical regions via standardized low-resolution 2D and 3D images. MedMNIST has rapidly become a popular testbed for DL architectures and automated model design in the biomedical domain. From a radiomics perspective, such a controlled benchmark is

* Corresponding authors.
Email addresses: alessandra.pernician@unica.it (A. Perniciano), federicom.cau@unica.it (F.M. Cau), luciod.spano@unica.it (L.D. Spano), cecilia.dir@unica.it (C. Di Ruberto), andrea.loddo@unica.it (A. Loddo).

https://doi.org/10.1016/j.neucom.2026.133854
Received 27 December 2025; Received in revised form 10 April 2026; Accepted 4 May 2026
Available online 5 May 2026

valuable because standardized preprocessing and fixed splits reduce confounding factors and facilitate reproducible, coherent comparisons of feature extraction and feature selection strategies across datasets and classifiers, enabling systematic analyses of redundancy, stability, and interpretability. Nonetheless, most existing studies focus on end-to-end DL models, with comparatively little attention to radiomics workflows based on handcrafted feature extraction and classical ML. As a result, the potential of MedMNIST as a controlled setting for investigating radiomics-specific issues, such as feature redundancy, feature selection strategies, and model interpretability, remains underexploited.

Feature selection (FS) is a key component of radiomic pipelines, aiming to identify compact, informative, and non-redundant feature subsets that preserve or improve predictive performance while enhancing robustness and interpretability. In a related work, it has been observed that FS is often treated as a secondary design choice, with limited effort devoted to aligning the FS strategy with the characteristics of both the dataset and the classifier [7]. Conventional FS techniques often fall short when applied to highly redundant radiomic spaces. Univariate filter methods, such as Analysis of Variance (ANOVA) or Mutual Information, evaluate features independently of the others. Consequently, they can present two important drawbacks: first, they fail to identify and remove highly correlated, redundant descriptors, and second, they neglect feature dependencies and model-specific interactions, which may discard variables that are weak individually but highly discriminative in combination. On the other hand, multivariate wrapper methods can capture these interactions but are often computationally prohibitive.

In recent years, eXplainable AI (XAI) has provided new tools for model-centric FS. Among these, SHapley Additive exPlanations (SHAP) [8] has attracted considerable interest due to its game-theoretic basis and its ability to quantify local and global feature contributions for complex models. SHAP-based FS has been shown to yield compact and competitive feature subsets across multiple domains [9–13], including biomedical applications [14,15].

While XAI tools like SHAP offer a model-aware alternative, applying SHAP directly to high-dimensional raw radiomic spaces imposes a substantial computational bottleneck. In our setting, direct SHAP-based elimination requires hundreds of additional model training/evaluation steps, and this cost can grow further with larger feature spaces. This motivates a pipeline that combines efficient redundancy reduction with model-aware evaluation.

More broadly, robustness-oriented representation design has been explored in other application domains beyond radiomics, including Face Anti-Spoofing (FAS) and distribution-shift settings in sequential data (e.g., cross-modal alignment for long-term network traffic forecasting). For instance, recent FAS work leverages Multimodal Large Language Models (MLLMs) to generate human-readable reasoning chains [16]. Similarly, CFPL-FAS [17] adopts causal intervention and feature-level constraints to promote invariance.

In sequential domains, CFAlignNet [18] employs knowledge-enhanced coarse-fine cross-modal alignment to handle distribution shifts and improve generalization. While the aforementioned studies employ different data modalities, feature representations, and objectives compared to handcrafted radiomics, they illustrate a broader methodological trend that promotes redundancy-aware and interpretable modeling considerations across multiple application areas.

In radiomics, SHAP has been employed both to interpret models and to drive FS [19–23]. However, on MedMNIST, most SHAP-based works concentrate on explaining DL representations or hybrid DL-XAI pipelines [24,25], rather than on SHAP-guided selection of handcrafted radiomic features. In addition, the computational cost of SHAP grows rapidly with the number of features and model evaluations [9,13], and the combination of correlation-based redundancy reduction with SHAP-based ranking has been only marginally explored in radiomics.

Moreover, interoperability is becoming an important requirement for radiomics pipelines: FS strategies should be applicable across different imaging modalities, datasets, and classifiers, while producing interpretable feature sets that can be compared across studies. Handcrafted radiomic features extracted using standardized libraries such as PyRadiomics [26] are a natural basis for such interoperability, provided that downstream FS methods are explicitly designed to be both model-aware and broadly applicable.

Building on these observations, we propose a radiomics-oriented two-stage feature selection pipeline, termed **Redundancy-Reduced SHAP-based Evaluation** (ReSHAPe), that integrates a correlation-based redundancy-reduction step with SHAP-driven ranking of radiomic features. In the first stage, highly correlated features are filtered out to mitigate redundancy; in the second stage, SHAP-based global importances computed on the training set guide the selection of a compact subset of relevant and interpretable features. We also investigate an ensemble variant that aggregates SHAP rankings across multiple classifiers.

To the best of our knowledge, our study is among the first to systematically evaluate a pipeline combining correlation-based filtering with SHAP-based feature ranking for handcrafted radiomics on the MedMNIST datasets.

We evaluate this pipeline on three MedMNIST v2 sub-datasets [2], considering both radiomic and non-radiomic biomedical settings, and compare it with baseline models trained on the complete feature set as well as with standard filter-based FS methods. The results show that, on radiomic datasets, *ReSHAPe* achieves performance comparable to or better than the baseline with only a small fraction of the original features and generally outperforms traditional univariate FS approaches. At the same time, pre-filtering for correlation substantially reduces the computational burden of SHAP, making SHAP-based FS more practical in radiomic workflows.

In summary, this work makes the following main contributions:

- We discuss a recurring limitation in radiomics practice, where FS is often chosen without a principled alignment to classifier and dataset characteristics, especially in benchmark settings such as MedMNIST [1,2,7].
- We introduce *ReSHAPe*, a two-stage FS pipeline that combines correlation-based redundancy reduction with SHAP-based global feature importance, explicitly tailored to the redundancy and interpretability requirements of radiomic features.
- We propose an interoperable evaluation setting in which handcrafted radiomic features and the proposed FS strategies are tested across multiple MedMNIST v2 datasets and ML classifiers, including an ensemble SHAP variant that promotes consensus feature subsets across models.
- We provide an experimental analysis demonstrating that our method can retain or improve predictive performance while drastically reducing feature dimensionality, thereby enhancing interpretability and computational efficiency within a standard radiomics workflow.

The remainder of this paper is organized as follows. In Section 2, we review MedMNIST-based benchmarking efforts and existing XAI-driven FS approaches. In Section 3, we introduce the proposed approach and detail how it integrates into the broader radiomics workflow, while Section 4 details the datasets, feature extraction procedures, classifiers, and evaluation metrics adopted in this study. The experimental setup and results are presented in Section 5, while Section 6 discusses the implications of our findings. Finally, Section 7 concludes the work and outlines directions for future research.

2. Related work

2.1. On the MedMNIST dataset

Classic datasets (MNIST, ImageNet) advanced generic machine learning, but are limited for medical imaging [1]. Biomedical options like CheXpert and BreakHis introduce disease-oriented tasks, but with high-resolution images and challenging access [2]. MedMNIST was curated

Table 1
Extended comparison of major benchmarks on MedMNISTv2, with metrics and best results provided by the referenced source works.

Paper & Citation	Models Evaluated	Task/Sub-datasets	Main Metrics	Best Reported Results	Contributions
Yang et al. (2023) [2]	Baseline CNNs, VGG, ResNet, DenseNet	All 18 sub-datasets (2D & 3D)	Accuracy, AUC, F1, Sens., Spec.	ResNet (OrganMNIST-A) avg. AUC ≈ 0.935; CNN/DenseNet strong on 2D	Introduced MedMNIST v2; standardized protocol; first full benchmark
Yang, Shi & Ni (2021) [1]	Classical ML, CNNs, AutoML (AutoKeras, AutoGluon)	Decathlon (10 sub-datasets)	Accuracy, AUC	CNN (PathMNIST) AUC ≈ 0.969; AutoML close to DL on some tasks	Original MedMNIST benchmark; proved AutoML feasibility
Zheng et al. (2024) [27]	NAS, AutoML, CNNs	PathMNIST, OrganMNIST, others	Accuracy, AUC	NAS pipelines PathMNIST AUC > 0.97	SOTA NAS/AutoML on MedMNIST
Halder et al. (2024) [28]	Vision Transformer (ViT), ResNet, CNN	BloodMNIST, BreastMNIST, PathMNIST, RetinaMNIST	Accuracy, F1, Precision, Recall	ViT: BloodMNIST 97.9%, BreastMNIST 90.4%, PathMNIST 94.6%, RetinaMNIST 57%	First ViT benchmark on MedMNISTv2 subsets; ViT outperformed CNNs on several sub-tasks
Zhang et al. (2025) [29]	ResNet-50, ViT, SwinV2-Tiny, LeViT, CvT-13, DINOv2	RetinaMNIST; Diabetic Retinopathy (DR)	Accuracy, QWK, AUC, F1	CvT-13: QWK 0.84, AUC 0.93; hybrid models outperform	First systematic comparison of CNN, ViT, and hybrid models for DR on MedMNISTv2
W. Liu et al. (2025) [30]	SRE-CNN (Symmetric Rotation Equivariant CNNs), ResNet	MedMNISTv2 (16 tasks)	Accuracy, AUC	SRE-Convoutperfo rms ResNet on rotation-sensitive tasks	Proved rotation equivariance boosts MedMNISTv2 performance
W. Zhang et al. (2025) [31]	KAN-Integrated MedViT, CNN, ViT, Hybrid	MedViTV2 (MedMNIST2D tasks)	Accuracy, AUC	MedViTV2 achieves new SOTA on medical 2D benchmarks	Established KAN-integrated ViT as new SOTA for MedMNIST2D
Saha et al. (2024) [32]	Deep Fuzzy Rank-based Ensemble	DermaMNIST (MedMNISTv2)	Accuracy	Ensemble: 94.1 accuracy	Deep fuzzy ensemble for robust multi-class skin disease classification

to offer diversity, accessibility, and rapid prototyping in the biomedical domain [1,2], and is featured on major platforms [33,34].

Classical ML methods (SVMs, random forests) remain relevant for benchmarking and small data. DL (e.g., CNNs) excels in large-scale, complex scenarios [2]. Automated model selection, including Automated Machine Learning (AutoML) such as Neural Architecture Search (NAS), has enabled non-specialists to compete with expert designs [27]. Recent MedMNIST studies incorporate AutoML for comprehensive benchmarking [1].

Key publications present MedMNIST (and v2) as a benchmark suite for multi-modal, low-resolution biomedical images and encourage reproducibility, automated system benchmarking, and educational activities [1,2]. An example of MedMNISTv2 used as a benchmark is shown in Table 1. Preliminary experiments using classical, deep learning (DL), and AutoML methods have established MedMNIST as the de facto lightweight benchmark [33].

2.2. On the use of SHAP in the medical field

Over the last decade, feature selection processes have benefited from XAI techniques that adopt model-centric selection strategies. One such approach is SHAP [8], which uses game-theoretic principles to estimate a model's feature importances on the training or test set; these importances are then ranked to identify the features most influential for the target task. In particular, SHAP has been shown to select feature subsets that deliver performance comparable to or superior to those derived from other XAI approaches [9–13] and has been employed to improve model performance while reducing feature dimensionality across domains such as driver state monitoring [35], financial forecasting [36], intrusion detection [37,38], and biomedicine [14,15].

This also applies to the radiomics domain, where SHAP was used both as an interpretability tool [39–41] and as a feature selection technique [19–23]. For example, Kha et al. [19] trained an XGBoost model to predict low-grade gliomas (LGGs) in "The Cancer Imaging Archive" (TCIA) public dataset, using SHAP as a feature selection method to identify the seven most informative image-derived features, which resulted in improved accuracy compared to considering the complete set of features.

Additionally, Goktas et al. [20] proposed an interpretable machine learning framework for cell image classification based on a dataset of human "Mesenchymal Stem Cells" (MSCs), using SHAP to improve model transparency and precision in cell characterization by identifying the most informative image-derived patterns after noise reduction. Recently, Samara et al. [22] analyzed "The Cancer Genome Atlas" (TCGA) glioma dataset, using SHAP as part of a feature selection pipeline to identify 13 key genetic and clinical biomarkers. The selected features were used to train Random Forest, Support Vector Machine, XGBoost, and Logistic Regression models, leading to improved interpretability and classification performance. Despite these advancements, few studies have used SHAP for feature selection on the MedMNIST dataset, focusing instead on deep learning approaches to extract image features [24,25]. In parallel, using correlation-based filtering (e.g., Pearson or Spearman [42,43]) as a preliminary step before SHAP-based ranking remains understudied in the radiomics domain, despite the well-documented computational cost of SHAP [9,13] and the potential of such sequential pipelines to mitigate this limitation.

Building on these premises, we propose a feature selection pipeline that combines Spearman correlation, accounting for feature distribution skewness, with SHAP global importance computed on the training set to guide the selection of radiomics features for multiple machine learning models.

3. ReSHAPe: two-stage redundancy-reduced SHAP feature selection

This work arises from the need for a feature selection approach that is naturally interpretable and flexible enough to accommodate various classifiers and datasets. This motivation stems from related work in which it has been highlighted that there is a lack of attention to selecting an appropriate feature selection method tailored to the specific classifier and dataset characteristics [7]. *ReSHAPe* was developed with consideration of the intrinsic characteristics of radiomic features, which are inherently redundant and therefore mutually correlated, as well as the strong relationship between radiomic features and the underlying biology. This association may hold significance only for a specific case [3–5].

ReSHAPe consists of two main steps: the first involves reducing features based on correlation, as described in Section 3.1, and its output serves as the input for the SHAP-based feature selection process detailed in Section 3.2.

3.1. Stage 1 – spearman-based redundancy pruning with skewness-based retention: dealing with correlation

It is well established in the literature that radiomic features often exhibit redundancy [6]. To address this limitation, we introduced a preprocessing step to reduce redundancy by removing features with high correlation [44]. Specifically, we computed the correlation matrix using Spearman's rank coefficients, which account for potential non-linear dependencies. From this matrix, only the upper triangular portion was considered, with diagonal elements set to zero to avoid self-comparisons. For each feature, we then recorded its set of correlated counterparts. Features were ranked by the number of correlations they exhibited, and pruning began with those exhibiting the highest degree of redundancy. For each correlated pair, we retained the feature with the most favorable distribution, defined as the feature with the absolute skewness closest to zero. In cases where a feature was correlated with multiple others, it was preserved only if its absolute skewness was lower than that of all its correlated counterparts; otherwise, it was discarded.

Choice of the correlation threshold. We define two features as redundant when their absolute Spearman rank correlation satisfies $|\rho_s| \geq \tau$. We set $\tau = 0.7$ to target *strong* monotonic dependencies, which are typically indicative of near-interchangeable radiomic descriptors, while avoiding overly aggressive pruning that may remove complementary information. Lower thresholds increase feature removal and reduce the computational burden of the subsequent SHAP stage, but may discard partially redundant yet informative descriptors; higher thresholds reduce pruning and retain a richer candidate set, but substantially increase SHAP overhead.

On τ as a design choice. We treat τ as an explicit *design parameter* controlling redundancy tolerance and computational budget, rather than as an opaque hyperparameter tuned to maximize test performance. In particular, τ has a direct and interpretable meaning (what level of dependence is deemed redundancy) and affects the number of features entering Stage 2, while keeping all retained descriptors in the original handcrafted radiomic space.

Rationale for skewness-based retention and empirical validation. We use absolute skewness as a criterion to choose a representative within highly correlated radiomic feature groups. Radiomic descriptors often exhibit asymmetric, heavy-tailed distributions due to quantization, Region Of Interest (ROI) heterogeneity, and effects from texture operators. When two features are nearly redundant (high $|\rho_s|$), retaining the feature with the absolute skewness closer to zero serves as a proxy for selecting a more symmetric, and typically more numerically stable, variable, reducing sensitivity to extreme tails after z-score normalization and improving robustness across models. To empirically validate this design choice, we compared our label-free retention strategy against a label-aware pruning strategy. Instead of retaining the feature with the lowest absolute skewness, we retained the feature with the highest Mutual Information (MI) with the target label (computed solely on the training set). Our experiments revealed that the MI-based approach did not consistently yield improvement.

All the results can be found in Table 2. These results support the design of our pipeline. Stage 1 is designed to be an unsupervised redundancy-reduction step, aimed purely at addressing multicollinearity by retaining the most numerically stable variable within a correlated cluster. Replacing this with an MI-based criterion would introduce a univariate supervised filter. Univariate filters evaluate features in isolation and may prematurely discard variables that exhibit low individual correlation with the target but possess high predictive value through multivariate interactions. By employing a label-agnostic metric, such as skewness, for redundancy reduction, ReSHAPe avoids this pitfall, delegating the complex task of evaluating task-specific relevance and feature interactions to the subsequent, model-aware SHAP selection stage.

Table 2
Classification results (weighted F-measure) comparing the label-free (Skewness) and label-aware (Mutual Information, MI) retention criteria during the correlation-based redundancy reduction stage.

Classifiers	Datasets					
	BreastMNIST		*PneumoniaMNIST*		*BloodMNIST*	
	Skewness	MI	Skewness	MI	Skewness	MI
Decision Tree	0.7140	0.7719	0.8885	0.8928	0.7713	0.7818
Random Forest	0.8817	0.8914	0.9576	0.9576	0.8737	0.8736
ExtraTree	0.9039	0.8887	0.9460	0.9595	0.8589	0.8668
XGBoost	0.8643	0.8937	0.9635	0.9596	0.9162	0.9077
SVM Linear	0.8436	0.8406	0.9750	0.9693	0.8889	0.8940

Statistical rationale for spearman correlation and skewness-based retention. Spearman's rank correlation coefficient ρ_s is defined as the Pearson correlation of the rank-transformed variables:

$$\rho_s(X, Y) = \frac{\mathrm{Cov}(\mathrm{rk}(X), \mathrm{rk}(Y))}{\sigma_{\mathrm{rk}(X)} \cdot \sigma_{\mathrm{rk}(Y)}} \tag{1}$$

This rank transformation confers two properties that are particularly desirable in radiomic feature spaces. First, it captures *monotonic* dependencies rather than only linear ones, which is important given that many radiomic descriptors are related through nonlinear transformations (e.g., energy is proportional to the square of pixel intensities, which are themselves correlated with percentile-based features). Second, the rank transformation is *robust to outliers and heavy-tailed distributions*: radiomic features extracted from heterogeneous regions of interest frequently exhibit skewed, non-Gaussian distributions, as confirmed empirically by the skewness values observed across our feature sets, making a rank-based measure statistically more appropriate than Pearson's r, which assumes approximate normality and is sensitive to extreme values. Indeed, the very presence of non-negligible skewness in the feature distributions, which motivates the retention criterion described below, retroactively justifies the choice of Spearman over Pearson.

Within a group of highly correlated features ($|\rho_s| \geq \tau$), all members carry approximately equivalent information about the target variable. The selection of a representative within each group is therefore a problem of *numerical stability* rather than one of information content. Formally, the skewness of a feature X with finite third central moment μ_3 and standard deviation σ is defined as:

$$\gamma_1(X) = \frac{\mu_3}{\sigma^3} \tag{2}$$

After z-score normalization, a feature with high absolute skewness $|\gamma_1|$ retains a long tail in the standardized domain. This concentrates statistical influence in a small number of samples, which can destabilize coefficient estimates in linear models, increase variance in tree-based split decisions, and bias SHAP value computations, all of which operate downstream in Stage 2. Retaining the feature with $|\gamma_1|$ closest to zero within each correlated cluster is therefore a principled, label-agnostic proxy for selecting the most numerically stable and statistically well-behaved representative. This theoretical justification is empirically supported by Table 2, which shows that replacing skewness-based retention with a label-aware MI criterion does not yield consistent improvements across datasets and classifiers, confirming that within tightly correlated groups, predictive performance is largely insensitive to the choice of retention criterion, making a numerically motivated, label-free criterion preferable.

Finally, we note a formal structural reason to avoid projection-based dimensionality reduction methods such as manifold learning in this

Algorithm 1 Correlation-based Feature Reduction.

```
Compute the Spearman correlation matrix C
Keep only the upper triangular part of C and set the diagonal to 0
Extract all feature pairs (f_i, f_j) with |C_ij| ≥ 0.7
Build a data structure M mapping each feature f_i to its correlated features
Count the number of correlations for each feature and sort features in descending order
for each feature f in descending order of correlation do
    Let G be the list of features correlated with f
    Compute absolute skewness for f and for all g ∈ G
    if |skew(f)| > |skew(g)| for any g ∈ G then
        Remove f and delete it from all lists in M
        break
    else if |skew(f)| <= |skew(g)| for all g ∈ G then
        Remove all features in G and update M
    end if
end for
```

context. Any embedding $\phi: \mathbb{R}^d \to \mathbb{R}^k$ that optimizes a geometric criterion (e.g., preservation of local neighborhoods in Uniform Manifold Approximation and Projection (UMAP), or geodesic distances in Isomap) produces latent components that are nonlinear combinations of all original features with no closed-form inverse. These components, therefore, carry no direct physical interpretation – a property that is not merely aesthetically undesirable in radiomics, but practically incompatible with clinical auditability requirements. Our Stage 1, by contrast, applies a binary selection mask entirely within the original feature space $\mathbb{R}^d$, ensuring that every retained descriptor preserves its original physical meaning (e.g., *90th percentile of pixel intensities*, *surface area*).

3.2. Stage 2 – SHAP as feature selection method (SHAPfs)

We employed SHAP-based feature selection (SHAPfs) to guide the selection of the most relevant features for the task. Specifically, the model was first trained on the complete set of features, and SHAP values were then computed. Since SHAP provides feature contributions for each class, we aggregated them by averaging across classes to obtain a single importance score for each feature. In particular, SHAP returns class-conditional attributions $\phi_{i,j,c}$ for sample i, feature j, and class c. Therefore, the global importance score across samples and then across classes was defined as follows:

$$I_j = \frac{1}{C}\sum_{c=1}^{C}\left(\frac{1}{S}\sum_{i=1}^{S}|\phi_{i,j,c}|\right), \tag{3}$$

For binary classification, the same procedure applies with $C = 2$ (equivalently, using only the positive class, which yields the same ranking up to a constant factor). After calculating the importance scores, features were ranked in descending order of importance, and the least relevant features were iteratively removed. This elimination process was repeated until the desired number of features was retained.

SHAPfs ensemble represents an extension of the previously described SHAP-based approach. In this method, feature selection is performed by first identifying the top 25 features for each classifier using SHAPfs. Following this, we select the 25 features that rank highest across all classifiers, establishing a common foundation for the next stage of our analysis.

3.3. End-to-end pipeline

This study follows the conventional radiomic workflow, encompassing image acquisition, feature extraction and normalization, feature selection, and model development and evaluation. Within this framework, we propose a novel hybrid feature selection strategy, which is structured as follows:

1. **Correlation-based filtering**: Initially, highly correlated features are removed following the algorithm described in Section 3.1, to mitigate redundancy and multicollinearity within the feature set.
2. **Model training**: The predictive model is subsequently trained using the filtered set of features as input.
3. **SHAPfs** (Section 3.2): The SHAP algorithm is then applied to the trained model to estimate the contribution of each feature to the model's output. Features are ranked by SHAP importance, and the least relevant feature is removed.

Steps 2 and 3 are repeated until a target budget of $k = 25$ features is reached. We fix $k = 25$ to obtain a compact and inspectable feature panel (25/285 descriptors, i.e., , $\approx 9\%$ of the original feature space), consistent with the goal of interpretability in radiomics and with the feature-budget analysis adopted in our experiments (we consider $k \in \{200, 100, 50, 25\}$ for standard filter baselines). **On k as a design choice.** We treat k as a *design parameter* that operationalizes the desired level of compactness and interpretability of the final radiomic signature, rather than as an opaque tuning knob. Smaller values of k yield more concise and easier-to-review feature sets, whereas larger values may preserve additional fine-grained information at the cost of reduced compactness. In this study, fixing $k = 25$ provides a strict and transparent feature budget that enables direct comparisons across classifiers and datasets under a highly interpretable regime.

All feature selection operations (correlation pruning and SHAP-based ranking and removal) are performed using the **training split only**; the validation/test splits are never used to compute correlations, SHAP importances, or normalization parameters.

The visual pipeline is shown in Fig. 1.

4. Materials and methods

This section describes the materials and methods employed in this study, including the datasets utilized (Section 4.1), the features extracted from these datasets (Section 4.2), the classification and feature selection methods used to evaluate and compare our proposed approach (Section 4.3), and the metrics applied for performance evaluation (Section 4.4).

4.1. Dataset: MedMNIST

MedMNIST v2 includes 18 standardized biomedical image subdatasets (12 for 2D and 6 for 3D tasks), carefully pre-processed for consistency and easy access [2]. Modalities include X-ray, pathology, endoscopy, and retinal images [2]. Each subset is divided into training, validation, and test folders with pre-determined splits. Images are resized to 28×28 pixels (2D) or $28\times28\times28$ voxels (3D). Further normalization or augmentation (rotation, flipping, contrast, etc.) is performed per model requirements [2]. The authors have released a new version, MedMNIST++, with greater sizes: 64×64 128×128, and 224×224 pixels for 2D, and $64\times64\times64$ voxels for 3D. All 2D images used in this work are 224×224 pixels.

The datasets used in this work are: **BreastMNIST**, **PneumoniaMNIST**, **BloodMNIST**. Specifically, *BreastMNIST* and *PneumoniaMNIST* are radiomic datasets, whereas *BloodMNIST* is not. Our decision to include a biomedical dataset that is not strictly radiomic was motivated by the aim of assessing whether the proposed method can also perform effectively when applied to radiomic features extracted from a non-radiomic dataset. In addition, BreastMNIST and PneumoniaMNIST are binary classification tasks, whereas BloodMNIST is a multiclass task.

BreastMNIST (Breast) is composed of 780 breast ultrasound images, split with a ratio of 7:1:2 into training, validation, and test sets. The classes were initially three, representing the outcomes of breast cancer: benign, malignant, and normal; however, since benign and normal are combined, the dataset is binary.

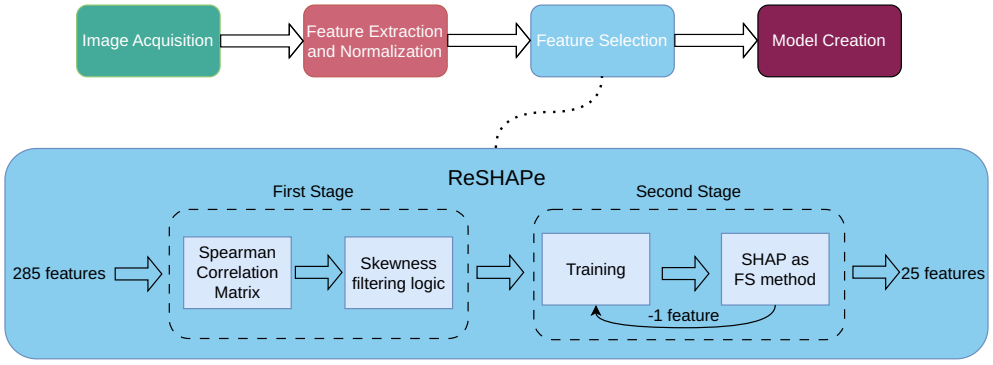


Fig. 1. We present the overall radiomics workflow, with a specific focus on the ReSHAPe pipeline. As illustrated, ReSHAPe serves as the dedicated feature selection module, strategically positioned after image acquisition, feature extraction, and data normalization. The ReSHAPe methodology consists of two distinct stages. The first is an unsupervised redundancy-reduction step that utilizes Spearman correlation to identify and filter out highly correlated features; when dealing with redundant pairs, the algorithm retains the feature exhibiting an absolute skewness closer to zero. The second stage is model-aware, employing SHAP techniques to evaluate global feature importance and iteratively discard the least relevant descriptors from the remaining feature space.

PneumoniaMNIST (Pneumonia) is a dataset of chest radiographs representing pediatric pneumonia. The task is binary, indicating whether pneumonia is present. The training set comprises 80% of all images, whereas only 20% are used for validation and testing, totaling 5856.
BloodMNIST (Blood) is composed of 17,092 single-cell images free of infection, hematologic, or oncologic illnesses. To differentiate among the many types of white blood cells, this dataset is organized into 8 classes. The ratio is also 7:1:2 in this case.

4.2. Feature extraction

The heart of Radiomics is the extraction of high-dimensional feature data to describe attributes of the region of interest quantitatively [3]. In this study, we extracted 285 radiomic features from the categories described in this section.
First-Order Features (Histogram Features). A color histogram was used to characterize the overall intensity distribution in the image. From this histogram, we extracted several statistical descriptors using both *MATLAB*[1] and the *PyRadiomics* library [26], including mean, standard deviation, smoothness, skewness, and kurtosis.
Second-Order Features (Texture Features). We focused on Haar-like features, first introduced by Viola et al. [45]. These features, which can be edge, line, four-rectangle, or center-surround, are composed of neighboring rectangles with alternating positive and negative polarity. An integral image, which enables quick computation of pixel sums within rectangular regions, is frequently used to compute Haar features.

Furthermore, we extracted thirteen Haralick features into rotation-invariant features called HAR_ri by extracting them from the Gray Level Co-Occurrence Matrix (GLCM) [46]. In particular, four different types of GLCM with $d = 1$ and $\theta = [0, 45, 90, 135]$ were calculated.

Finally, we extracted texture-related features from the GLCM using the *pyradiomics* library. We used the default settings of *pyradiomics* for the extraction. Among these were, for instance, *autocorrelation*, which quantifies the degree of texture fineness or coarseness, and the *difference average*, which characterizes the balance between occurrences of voxel pairs with similar versus differing intensity values.

Some features were also extracted from the following matrix:

- **Gray Level Size Zone Matrix (GLSZM)**[47]: quantifies gray-level zones within an image, where a zone is defined as the number of connected pixels sharing the same gray-level intensity.
- **Gray Level Run Length Matrix (GLRLM)**[48]: quantifies gray-level runs, defined as the length (in pixels) of consecutive pixels with the same gray-level intensity along a specific direction.
- **Neighboring Gray Tone Difference Matrix (NGTDM)**[49]: quantifies the difference between the gray value of a pixel and the average gray value of its neighbors within a distance δ; the sum of absolute differences for each gray level i is stored in the matrix.
- **Gray Level Dependence Matrix (GLDM)**[50]: quantifies gray-level dependencies in an image, where a dependency is defined as the number of connected pixels within a distance δ that are dependent on the central pixel.

We also used the feature vector obtained from the Local Binary Pattern (LBP) histogram in its rotation-invariant form [51].
Invariant Moments. A way to extract some features is to use the weighted average of pixel intensities in an image, known as the moment. In this work, we utilized three types of moments:

- *Chebyshev* moments (CH) [52]: Discrete orthogonal moments are implemented without requiring numerical approximation. Chebyshev polynomials serve as the basis for these moments, allowing the extraction of global image features by adjusting the moment order [53]. In our work, we used first-order and second-order Chebyshev moments, with vector orders of 5 and 4, respectively.
- *Legendre* moments (LM): First introduced by Teague [54], are orthogonal moments constructed to minimize redundancy within a set of moment functions, thus emphasizing independent features [55]. In our study, we employed a continuous fifth-order vector of Legendre moments, approximated using Simpson's rule.
- *Zernike* moments (ZM): Derived from Zernike polynomials, an orthogonal set defined over the unit disk. They provide a compact, non-redundant representation of image features, making them widely useful in image analysis tasks [56]. In this work, we employed a Zernike vector of order 6 with 4 repetitions.

4.3. Classification and feature selection methods

We applied machine learning techniques to train classification models for the various tasks and employed several feature reduction strategies, including our proposed method combined with state-of-the-art feature selection methods.

4.3.1. Machine learning methods

In this study, we used the following machine learning methods:

- Support Vector Machines;
- Decision Tree;

[1] *MATLAB*, The MathWorks, Inc. Available at: https://www.mathworks.com/products/matlab.html (Accessed: 2025-12-26).

- Random Forest;
- Extra Trees;
- eXtreme Gradient Boosting.

Support Vector Machines (SVM) [57] are state-of-the-art classifiers capable of modeling complex decision boundaries and handling high-dimensional feature spaces. Linear SVM seeks an optimal hyperplane that maximizes the margin between classes. The soft margin formulation and the kernel trick allow SVM to handle non-linearly separable problems.

A **Decision Tree (DT)** [58] is a predictive modeling approach that links input features to a target variable by recursively partitioning the data according to feature values. This process creates a tree-like structure in which each branch, from the root to a leaf, corresponds to a sequence of decision rules that culminate in a predicted outcome.

Three ensemble classifiers were employed:

- **Random Forest (RF)** [59]: is an ensemble of decision trees constructed using bagging, i.e., resampling with replacement of the training data. Each tree outputs a class prediction, and the final prediction is determined by majority voting across the ensemble, improving performance over a single decision tree.
- **Extra Trees (ET)** [60]: is an ensemble method similar to Random Forest, but differs in two main ways: splits are selected completely at random without optimizing thresholds, and each tree typically uses the entire dataset rather than bootstrap samples. Predictions are made by majority voting, and extreme randomization enhances the robustness of the model.
- **eXtreme Gradient Boosting (XGBoost)** [61]: is a gradient boosting algorithm in which new trees are trained on the residuals of previous trees. A gradient descent algorithm minimizes the loss at each step, allowing each subsequent tree to correct the errors of its predecessors, resulting in an improved final prediction.

4.3.2. Feature selection methods

Not considering the one that we are proposing we also used two univariate filter methods only for comparison. **f_classif** is a univariate feature selection method that relies on Analysis of Variance (ANOVA) [62]. It applies an F-test to each feature individually to assess the degree to which the feature discriminates between different target classes. **mutual_info_classif** is the implementation provided by the scikit-learn library [62] for feature selection based on Mutual Information. This method returns an array of mutual information scores, each quantifying the degree of dependency between a given feature and the target variable.

4.4. Evaluation metrics

The evaluation was performed using the following metrics:

- *True Positive Rate (TPR)* [58], also referred to as *Sensitivity* or *Recall*, quantifies the proportion of positive instances correctly identified by the classifier:

$$TPR = \frac{TP}{TP + FN} \tag{4}$$

  where TP denotes the number of positive instances correctly classified and FN the number of positive instances misclassified as negative.
- *Precision* [58], measures the proportion of correctly predicted positive instances among all instances classified as positive:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

  where FP indicates the number of negative instances classified erroneously as positive.
- *F-measure (F1-score)* [58], represents the harmonic mean of precision and recall, providing a measure when the class distribution is unbalanced:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

- *Accuracy* [58], reflects the overall proportion of correctly classified instances across all classes:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

- *Area Under the ROC Curve (AUC)* [58], evaluates classifier performance based on the *Receiver Operating Characteristic (ROC)* curve, which illustrates the trade-off between TPR and the False Positive Rate (FPR) at varying decision thresholds. A higher AUC indicates that the classifier is closer to the ideal scenario, where $TPR = 1$ and $FPR = 0$.

The reported averages include the macro average, computed as the unweighted mean of the metric across all classes, and the weighted average, calculated as the mean across classes weighted by the number of instances (support) in each class.

4.5. Technological setup

For the feature extraction step, we employed the **PyRadiomics** library [26] in Python, alongside our custom framework implemented in **MATLAB** for computing invariant moments [53]. All the experiments were conducted on a workstation equipped with an Apple M3 Pro CPU with 18 GB of RAM.

5. Experiments

In this section, we describe the experimental setup (Section 5.1) and present the results obtained from these experiments (Section 5.2).

5.1. Experimental setup

In this study, we introduce a novel SHAP-based feature selection method designed to enhance explainability within a radiomic workflow. To contextualize its behavior, we compare the proposed approach against alternative methods that share similar properties. In particular, mutual information, which quantifies the dependence between pairs of variables and is often employed as an alternative to correlation, is selected as a natural comparator. Given that our method adopts correlation as a preliminary filtering criterion, mutual information provides an appropriate benchmark for assessing the impact of different dependency measures.

Moreover, while correlation characterizes both the direction and strength of linear associations between random variables, ANOVA evaluates whether statistically significant differences exist among the means of multiple groups. We therefore include *f_classif*, an ANOVA-based univariate feature selection procedure, to capture an additional, complementary notion of the relationship between features and class labels in our radiomic study.

All classifiers described in Section 4.3.1 were implemented using *Scikit-Learn* [62] with their default parameters.

For each classifier, we performed different configurations:

- **Baseline**;
- **Mutual Information (MutualInfo)**;
- **f_classif**;
- **Correlation (Corr)**;
- **SHAPfs**;
- **SHAPfs Ensemble (SHAPfs-Ens)**;
- **ReSHAPe**;
- **ReSHAPe-Ens**.

In the *Baseline* configuration, the classifier is trained on all 285 features with no parameter modifications. *MutualInfo* and *f_classif* are

applied as standard feature selection methods. We also applied the algorithm described in Section 3.1 (*Corr*), which resulted in different numbers of features for each dataset: 69 features for *Breast*, 55 features for *Pneumonia*, and 70 features for *Blood*. The *SHAPfs* and *SHAPfs-Ens* configurations correspond to the application of the algorithms described in Section 3.2. *ReSHAPe-Ens* is a modification of *ReSHAPe*, applied within *SHAPfs-Ens* rather than *SHAPfs*.

MutualInfo and *f_classif* were evaluated under feature budgets $k \in \{200, 100, 50, 25\}$. In our experiments, the most restrictive budget ($k = 25$) led to a marked performance drop compared with larger cut-offs. Accordingly, we emphasize $k = 25$ for our method to show that ReSHAPe remains competitive even under a highly constrained setting, while yielding a compact and interpretable feature panel (25/285 features, i.e., $\approx 9\%$).

After extracting the features, *z-score normalization* was applied using the implementation provided by the *Scikit-Learn* package [62]. This method standardizes features by subtracting the mean and scaling to unit variance, transforming the data so that each feature has a mean of 0 and a standard deviation of 1.

All experiments follow the official MedMNIST v2 pre-defined *train/validation/test* splits. To prevent information leakage, the **test split is never used** during feature selection, ranking, or normalization. Specifically, for each dataset, classifier, and configuration:

- **Normalization:** z-score parameters (mean and standard deviation) are estimated on the **training split only** and then applied unchanged to validation and test.
- **Correlation filtering (Corr/ReSHAPe stage 1):** the Spearman correlation matrix and the skewness statistics used to retain a representative feature within correlated groups are computed **only on the training split**. The resulting feature mask is then applied to validation and test without recomputing correlations.
- **SHAP-based selection (SHAPfs/ReSHAPe stage 2):** at each iteration of the elimination loop, the model is fit on the **training split** using the current feature set, and SHAP values are computed **on the training split only**; feature removal decisions are made exclusively from these training-derived SHAP importances. The same procedure is applied to the ensemble variants: SHAP rankings are computed independently per classifier on the training data and then aggregated.

Unless explicitly stated otherwise, the validation split is used only for intermediate reporting or cutoff comparison, while the final performance is reported on the held-out test split. We do not employ nested cross-validation, as the goal is to benchmark feature selection strategies under the standardized MedMNIST evaluation protocol.

5.1.1. SHAP configuration

To compute SHAP values, we used the SHAP library and selected the explainer according to the classifier family. Specifically, we used: (i) *TreeExplainer* for tree-based models (DT, RF, ET, XGBoost); and (ii) *LinearExplainer* for linear models when applicable (e.g., linear SVM). All SHAP computations were performed using the **training split only**.

5.2. Results

Performance metrics for the experiments are summarized in Table 3, with corresponding completion times provided in Table 4.

For simplicity, we report only the outcomes using 25 features across different cut-off configurations, with the weighted f_measure as the evaluation metric, as this offers a more representative overview of the study cases.

Breast: Overall, the results obtained with reduced dimensionality are consistent with those of the baseline, which uses all features. The traditional feature selection methods (*Mutual Info* and *f_classif*) show slightly lower performance.

Pneumonia: With the exception of the DT classifier, which demonstrates superior performance using the complete feature set, all other classifiers achieve comparable, and in some cases even improved, results while utilizing fewer than 91% of the features.

Blood: As observed, the baseline configuration outperforms the other approaches, with the exception of the SHAPfs Ensemble, which achieves comparable and occasionally superior results. Overall, the differences in performance are relatively small. It is important to note that Blood is the only non-radiomic dataset included in the study; nevertheless, our proposed approach demonstrates strong performance even in this context.

6. Discussion

In this section, we present a discussion of the experimental results and compare ReSHAPe with the most relevant configurations in Section 6.1; we discuss the time complexity in Section 6.2 and, finally, we present several insights on the actual interpretability using a supporting case study in Section 6.3.

Table 3
Classification results (weighted F-measure) for BreastMNIST, PneumoniaMNIST, and BloodMNIST using $k = 25$ features (all methods except baseline and Corr).

Classifiers	Baseline	MutualInfo	f_classif	Corr	SHAPfs	SHAPfs Ens.	ReSHAPe	ReSHAPe-Ens
BreastMNIST								
Decision Tree	0.7500	0.8320	0.7140	0.7140	0.5095	0.8141	0.7501	0.7179
Random Forest	0.8915	0.8524	0.8524	0.8817	0.8672	0.8915	0.9061	0.8792
ExtraTree	0.9186	0.8100	0.8141	0.9039	0.9061	0.8764	0.9080	0.9080
XGBoost	0.8938	0.8672	0.8256	0.8643	0.8643	0.8958	0.8837	0.9096
SVM Linear	0.7287	0.7425	0.7530	0.8436	0.8449	0.8672	0.8141	0.8524
PneumoniaMNIST								
Decision Tree	0.9243	0.8654	0.8678	0.8885	0.9131	0.9041	0.9026	0.9003
Random Forest	0.9577	0.9226	0.9124	0.9576	0.9558	0.9617	0.9633	0.9516
ExtraTree	0.9521	0.9298	0.9202	0.9460	0.9503	0.9597	0.9556	0.9634
XGBoost	0.9673	0.9112	0.9131	0.9635	0.9674	0.9597	0.9634	0.9615
SVM Linear	0.9692	0.9068	0.8963	0.9750	0.9750	0.9635	0.9674	0.9596
BloodMNIST								
Decision Tree	0.8025	0.7499	0.7332	0.7713	0.8178	0.8300	0.7842	0.7859
Random Forest	0.8974	0.8205	0.8164	0.8737	0.9036	0.9069	0.8795	0.8862
ExtraTree	0.8807	0.8184	0.8016	0.8589	0.8980	0.9116	0.8790	0.8808
XGBoost	0.9347	0.8607	0.8339	0.9162	0.9279	0.9264	0.9144	0.9059
SVM Linear	0.9397	0.8374	0.8158	0.8889	0.9120	0.9081	0.8793	0.8846

Table 4
Cumulative execution times (seconds) for filtering and classification on BreastMNIST, PneumoniaMNIST, and BloodMNIST ($k = 25$).

Classifiers	Baseline	MutualInfo	f_classif	Corr	SHAPfs	SHAPfs Ens.	ReSHAPe	ReSHAPe-Ens
BreastMNIST								
Decision Tree	0.1105	0.3933	0.0234	0.1560	20.0693	811.7682	3.5900	140.1020
Random Forest	0.3068	0.5051	0.1446	0.3135	180.3230	811.8834	31.2200	140.2171
ExtraTree	0.1167	0.4553	0.0965	0.2307	521.6429	811.8349	88.5165	140.1686
XGBoost	0.2311	0.4736	0.0908	0.2680	49.1383	811.8263	8.6019	140.1600
SVM Linear	0.2632	0.4344	0.0628	0.3239	49.5801	811.7999	8.6781	140.1337
PneumoniaMNIST								
Decision Tree	0.7228	2.9659	0.1586	0.5117	160.4358	24,977.1396	19.3940	781.6615
Random Forest	2.6051	3.6719	0.8682	1.8129	4916.1400	24,977.1396	584.2478	781.6615
ExtraTree	0.4468	3.2557	1.2462	0.8410	19,281.9010	24,977.1396	2290.5259	781.6615
XGBoost	1.5623	3.1350	0.4544	0.9268	282.9495	24,977.1396	33.9455	781.6615
SVM Linear	2.1688	3.8645	0.9725	1.4106	335.7133	24,977.1396	40.2124	781.6615
BloodMNIST								
Decision Tree	3.7849	9.3778	0.5448	1.5149	2042.9992	757,161.6425	360.7826	133,447.0729
Random Forest	10.6703	12.9117	4.2744	6.0546	178,775.1446	757,165.3605	31,508.9768	133,450.7909
ExtraTree	1.9086	10.2227	1.3518	2.2313	563,048.8142	757,162.4404	99,235.3706	133,447.8708
XGBoost	27.6429	14.1481	5.4490	10.9368	10,817.0771	757,166.5368	1907.1718	133,451.9671
SVM Linear	19.5775	17.5186	8.8323	15.7840	2477.3089	757,169.9322	437.3276	133,455.3626

6.1. Comparison between ReSHAPe and the other methods

ReSHAPe vs Baseline. Based on the results of our experiments, we conclude that across all datasets, our method does not lead to a marked decrease under the official fixed MedMNIST splits; observed differences are generally small. In particular, for the Breast and Pneumonia datasets, the proposed approach even outperforms the baseline. It is worth noting that our method achieves these results while utilizing fewer than 9% of the original features. Moreover, these selected features are not only relevant to the specific tasks but also interpretable, as they are entirely handcrafted.

ReSHAPe vs Corr. Although correlation pruning (Corr) alone already provides a strong reduction of redundancy, the addition of SHAPFs plays a complementary role by refining the reduced feature space in a model-aware manner under a fixed feature budget ($k = 25$). Table 3 highlights that ReSHAPe often improves upon Corr for tree-based models in both radiomic datasets. For example, on BreastMNIST, ReSHAPe yields higher weighted F1 than Corr for DT, RF, ET, and XGBoost, while showing a slight decrease for linear SVM. A similar pattern is observed on PneumoniaMNIST, where ReSHAPe improves upon Corr for DT, RF, and ET, remains essentially unchanged for XGBoost, and shows a small decrease for SVM.

On BloodMNIST (non-radiomic and multiclass), Corr remains competitive, and ReSHAPe provides mixed outcomes: it improves Corr for DT, RF, and ET, while slightly underperforming Corr for XGBoost and SVM. This suggests that in multiclass settings with fine-grained distinctions, correlation-based pruning may occasionally remove partially redundant yet complementary descriptors that can still be useful for certain classifiers, whereas the SHAP stage tends to favor a compact set optimized for the trained model under the strict $k = 25$ constraint.

Overall, these observations indicate that the first stage of ReSHAPe reduces multicollinearity and SHAP computational overhead, whereas the second phase contributes an additional model-aware refinement that can recover or improve performance under a compact and interpretable signature budget, with some classifier and dataset dependent variability under fixed splits.

ReSHAPe vs SHAPfs. In our study, we investigated whether using SHAP as a feature selection method, without prior filtering based on feature correlation, could yield better performance. For the Breast and Pneumonia datasets, SHAPfs outperforms our approach in only 3 of 10 cases, with the largest observed performance difference being 0.03. A different trend is observed with the Blood, where the performance of *ReSHAPe* is consistently lower than that of SHAPfs, with a maximum decrease of 0.04. This is likely due to the specific nature of Blood: being a non-radiomic, multiclass problem, the handcrafted radiomic descriptors extracted from cell images may benefit from retaining partially redundant but complementary features, so that the correlation-based pruning step in *ReSHAPe* can discard variables that are weakly correlated yet still discriminative for fine-grained class distinctions. In this setting, applying SHAP directly to the whole feature space enables the model to leverage a richer set of non-correlated features, resulting in slightly improved performance at the expense of higher computational complexity.

In general, it is important to note that applying SHAP to the full set of 285 features requires approximately 200 additional training steps, making it highly computationally intensive.

ReSHAPe vs ReSHAPe-Ens By applying the ensemble paradigm, we initially expected an overall improvement in performance; however, we observed an increase only for the Blood subset. This outcome may be attributed to the complexity of the dataset, which is multiclass (seven classes) and exhibits class imbalance. Consequently, the classifier may benefit from features generalized across multiple models rather than relying solely on those derived from a single classifier. In contrast, the Breast and Pneumonia datasets may not require features deemed necessary by other classifiers to achieve optimal performance.

Mutual Info vs Corr. *Corr* consistently outperforms *Mutual Info* for Pneumonia and Blood, whereas for Breast, *Mutual Info* performs better in only two out of five cases. By definition, Mutual Information involves a discretization step to compute joint probabilities; consequently, features that exhibit even minimal variations, such as invariant moments, may lose representativeness during this process, thereby reducing overall performance.

6.2. Time complexity discussion

Table 4 details the execution time for each method, along with their respective classification performance. The data show a direct correlation between method complexity and computational overhead. Notably, *ReSHAPe* emerges as the most efficient among the proposed SHAP-based techniques; its preliminary filtering phase effectively eliminates redundant features via correlation analysis, thereby streamlining the computational process. It is essential to clarify that the proposed methodology is intended for use during the training phase to identify and generalize subsets of relevant features. Consequently, these features are used directly during inference, ensuring that the associated computational complexity is strictly limited to the offline training phase and does not affect real-time performance.

6.3. On the interpretability of selected features: a case study on BreastMNIST

To further illustrate the interpretability of the redundancy-reduction stage, we examine concrete examples of intra-class and inter-class feature correlation in the context of BreastMNIST, grounding them in the physical properties of breast ultrasound imaging. Breast ultrasound images frequently exhibit high-intensity regions arising from hyperechoic structures, i.e., areas that strongly reflect ultrasound waves and therefore appear brighter [63]. This phenomenon is driven by the composition of breast tissue: glandular and fibrous tissue tends to appear in white or light-gray tones, while Cooper's ligaments manifest as bright white lines or streaks traversing the adipose tissue [64,65]. These physical characteristics have a direct impact on the statistical distribution of pixel intensities, and consequently on the radiomic features extracted from these images.

Intra-class redundancy. Within the first-order feature family, entropy, the 90th percentile, and energy exhibit strong mutual correlation. Entropy quantifies the randomness of the pixel intensity distribution; the 90th percentile represents the intensity value below which 90% of pixels fall; and energy is defined as the sum of squared pixel intensities. When many high-intensity pixels are present, as is typical in breast ultrasound images rich in hyperechoic structures, the 90th percentile is elevated, thereby increasing energy. If these bright pixels are spatially scattered rather than clustered, entropy also increases. Consequently, in images containing diffuse reflective structures such as glandular tissue, fibrous tissue, and Cooper's ligaments, all three features tend to rise together, making their joint inclusion in a feature set largely redundant.

Inter-class redundancy. Redundancy also arises across different feature families. Consider the Root Mean Square (RMS) intensity, the Small Dependence Low Gray Level Emphasis (SDLGLE) from the GLDM, and the High Gray Level Zone Emphasis (HGLZE) from the GLSZM. RMS captures the average signal power; SDLGLE weights small pixel clusters with low gray-level intensity (inversely related to overall brightness); and HGLZE assigns a higher weight to spatially connected zones with high gray-level values. In a globally bright image, the average signal power (RMS) increases, while HGLZE increases and SDLGLE decreases in a correlated fashion. Although these features originate from distinct mathematical frameworks—first-order statistics, gray-level dependence, and gray-level zone size—they collectively describe different facets of the same underlying phenomenon: the distribution of luminance across the image. Their partial redundancy justifies their joint pruning by Stage 1 of ReSHAPe. These examples underscore that the redundancy removed by Stage 1 is not arbitrary: it reflects physically meaningful co-variation among features rooted in the imaging modality itself. By retaining only the most numerically stable representative of each correlated group, ReSHAPe preserves interpretability while avoiding the confounding effect of near-duplicate descriptors in downstream model training.

7. Conclusion

In this work, we introduced a novel feature selection method, *ReSHAPe*, to address a gap in the current radiomics literature. Although feature selection is recognized as necessary, it is often underexplored. *ReSHAPe* specifically tackles the inherent redundancy of radiomic features without assuming a predefined relationship between the organ and the corresponding radiomic images. Indeed, since medical images are generated by the interaction of radiation or ultrasound with tissues and organs, they are not merely visual representations but also reflect the body's physical properties; consequently, the features extracted from them retain this underlying information. Through experiments on two radiomic and one biomedical dataset, we demonstrated that selecting the most relevant features can achieve comparable, and in some cases superior, performance while utilizing only 9% of the original feature set. Specifically, our pipeline achieves an approximately 91% reduction in feature dimensionality (retaining a highly interpretable subset of 25 out of 285 features). Despite this substantial reduction, ReSHAPe remains highly competitive with the full-feature baselines and often matches or exceeds conventional univariate filters such as Mutual Information and ANOVA, with the largest gains reaching roughly 5–10 percentage points in some settings. As future work, we aim to mitigate the computational overhead of the training phase by investigating optimization techniques and exploring parallel processing frameworks to improve the method's scalability. In addition, we plan to conduct a stability analysis to assess the robustness of *ReSHAPe* and to extend the experimental evaluation to additional non-radiomic case studies to further validate the generalizability of the proposed method in the biomedical field.

CRediT authorship contribution statement

Alessandra Perniciano: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Federico Maria Cau:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Formal analysis, Conceptualization. **Lucio Davide Spano:** Writing – review & editing, Validation, Supervision. **Cecilia Di Ruberto:** Writing – review & editing, Validation, Supervision. **Andrea Loddo:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Given his role as Editor, he had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. - A.L. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5 - Call for tender No. 3277 published on December 30, 2021 by the Italian Ministry of University and Research (MUR) funded by the European Union - NextGenerationEU. Project Code ECS0000038 - Project Title eINS Ecosystem of Innovation for Next Generation Sardinia - CUP F53C22000430001- Grant Assignment Decree No. 1056 adopted on June 23, 2022 by the Italian Ministry of University and Research (MUR), and by the GNCS Project 2025 "Metodi di approssimazione globale per operatori integrali e applicazioni alle equazioni funzionali" (CUP E53C24001950001).

Data availability

The data are publicly available and we cited the owners.

References

[1] J. Yang, R. Shi, B. Ni, MedMNIST classification decathlon: a lightweight AutoML benchmark for medical image analysis, 2021, arXiv preprint arXiv:2010.14925. https://arxiv.org/abs/2010.14925.

[2] J. Yang, R. Shi, Z. Liu, D. Wei, B. Ke, L. Zhao, Z. Huang, X. Li, B. Ni, MedMNIST v2: a large-scale lightweight benchmark for 2D and 3D biomedical image classification, Sci. Data 10 (1) (2023) 1–14, https://doi.org/10.1038/s41597-022-01721-8, https://www.nature.com/articles/s41597-022-01721-8.

[3] C. McCague, S. Ramlee, M. Reinius, I. Selby, D. Hulse, P. Piyatissa, V. Bura, M. Crispin-Ortuzar, E. Sala, R. Woitek, Introduction to radiomics for a clinical audience, Clinical Radiology 78 (2) (2023) 83–98, https://doi.org/10.1016/j.crad.2022.08.149

[4] R.J. Gillies, P.E. Kinahan, H. Hricak, Radiomics: images are more than pictures, they are data, Radiology 278 (2) (2016) 563–577, https://doi.org/10.1148/radiol.2015151169

[5] C. Scapicchio, M. Gabelloni, A. Barucci, D. Cioni, L. Saba, E. Neri, A deep look into radiomics, La radiologia medica 126 (10) (2021) 1296–1311, https://doi.org/10.1007/s11547-021-01389-x

[6] M.E. Mayerhoefer, A. Materka, G. Langs, I. Häggström, P. Szczypiński, P. Gibbs, G. Cook, Introduction to radiomics, J. Nucl. Med. 61 (4) (2020) 488–495, https://doi.org/10.2967/jnumed.118.222893

[7] A. Perniciano, A. Loddo, C. Di Ruberto, B. Pes, Insights into radiomics: impact of feature selection and classification, Multimed. Tools Appl. 84 (26) (2024) 31695–31721, https://doi.org/10.1007/s11042-024-20388-4

[8] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 4768–4777, NIPS'17.

[9] W.E. Marcilio, D.M. Eler, From explanations to feature selection: assessing SHAP values as feature selection mechanism, 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI) (2020) 340–347, https://api.semanticscholar.org/CorpusID:227221383.

[10] A. Gramegna, P. Giudici, SHAP and LIME: an evaluation of discriminative power in credit risk, Front. Artif. Intell. 4 (Sep 2021), https://doi.org/10.3389/frai.2021.752558

[11] Y. Shi, Y. Zou, J. Liu, Y. Wang, Y. Chen, F. Sun, Z. Yang, G. Cui, X. Zhu, X. Cui, F. Liu, Ultrasound-based radiomics XGBoost model to assess the risk of central cervical lymph node metastasis in patients with papillary thyroid carcinoma: individual application of SHAP, Front. Oncol. 12 (Aug 2022), https://doi.org/10.3389/fonc.2022.897596

[12] Y. Gebreyesus, D. Dalton, S. Nixon, D. De Chiara, M. Chinnici, Machine learning for data center optimizations: feature selection using Shapley additive exPlanation (SHAP), Future Internet 15 (3) (2023), https://doi.org/10.3390/fi15030088, https://www.mdpi.com/1999-5903/15/3/88.

[13] H. Wang, Q. Liang, J.T. Hancock, T.M. Khoshgoftaar, Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods, J. Big Data 11 (1) (2024) 44, https://doi.org/10.1186/s40537-024-00905-w

[14] C. Sebastián, C.E. González-Guillén, A feature selection method based on Shapley values robust for concept shift in regression, Neural Comput. Appl. 36 (23) (2024) 14575–14597, https://doi.org/10.1007/s00521-024-09745-4

[15] I. Madakkatel, E. Hyppönen, LLpowershap: logistic loss-based automated Shapley values feature selection method, BMC Med. Res. Methodol. 24 (1) (2024) 247, https://doi.org/10.1186/s12874-024-02370-8

[16] G. Zhang, K. Wang, H. Yue, A. Liu, G. Zhang, K. Yao, E. Ding, J. Wang, Interpretable face anti-spoofing: enhancing generalization with multimodal large language models, 2025, arXiv preprint arXiv:2501.01720. https://arxiv.org/abs/2501.01720.

[17] A. Liu, S. Xue, J. Gan, J. Wan, Y. Liang, J. Deng, S. Escalera, Z. Lei, CFPL-FAS: class free prompt learning for generalizable face anti-spoofing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 222–232, https://doi.org/10.1109/CVPR52733.2024.00029

[18] A. Han, Z. Ye, Y. Zhou, X. Huang, CFAlignNet: knowledge-enhanced coarse-fine cross-modal alignment for long-term network traffic forecasting, Available at SSRN 5650682 (2025), https://doi.org/10.2139/ssrn.5650682

[19] Q.-H. Kha, V.-H. Le, T.N.K. Hung, N.Q.K. Le, Development and validation of an efficient MRI radiomics signature for improving the predictive performance of 1p/19q co-deletion in lower-grade gliomas, Cancers (Basel) 13 (21) (2021) 5398.

[20] P. Goktas, R. Simon Carbajo, PPSW–SHAP: towards interpretable cell classification using tree-based SHAP image decomposition and restoration for high-throughput bright-field imaging, Cells 12 (10) (2023), https://doi.org/10.3390/cells12101384, https://www.mdpi.com/2073-4409/12/10/1384.

[21] M. Monteleone, F. Camagni, S. Percio, L. Morelli, G. Baroni, S. Gennai, P. Govoni, C. Paganelli, Validating an explainable radiomics approach in non-small cell lung cancer combining high energy physics with clinical and biological analyses, Physica Medica 136 (2025) 105054, https://doi.org/10.1016/j.ejmp.2025.105054, https://www.sciencedirect.com/science/article/pii/S1120179725001644.

[22] M.N. Samara, K.D. Harry, Integrating boruta, LASSO, and SHAP for clinically interpretable glioma classification using machine learning, Biomedinformatics 5 (3) (2025), https://doi.org/10.3390/biomedinformatics5030034, https://www.mdpi.com/2673-7426/5/3/34.

[23] L. Chen, Z. He, Q. Ni, Q. Zhou, X. Long, W. Yan, Q. Sui, J. Liu, Dual-radiomics based on SHapley additive explanations for predicting hematologic toxicity in concurrent chemoradiotherapy patients, Discov. Oncol. 16 (1) (2025) 541, https://doi.org/10.1007/s12672-025-02336-2

[24] Y. Hu, A. Chaddad, SHAP-integrated convolutional diagnostic networks for feature-selective medical analysis, 2025, arXiv preprint arXiv:2503.08712. https://arxiv.org/abs/2503.08712.

[25] O.B. Guney, K.S. Saichandran, K. Elzokm, Z. Zhang, V.B. Kolachalama, Active feature acquisition via explainability-driven ranking, Proc. Mach. Learn. Res. 267 (2025) 20748–20765.

[26] J.J. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R.G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, H.J. Aerts, Computational radiomics system to decode the radiographic phenotype, Cancer Res. 77 (21) (2017) e104–e107, https://doi.org/10.1158/0008-5472.CAN-17-0339

[27] Y. Zheng, et al., Neural architecture search for biomedical image classification: a comparative study, J. Biomed. Inform. (2024), https://doi.org/10.1016/j.jbi.2024.104238, https://www.sciencedirect.com/science/article/abs/pii/S0933365724003063.

[28] S. Halder, K.T.A. Siddiqui, S. Shamim, ViT-MNIST: vision transformer benchmarks for multiple biomedical image classification, Sci. Rep. 14 (1) (2024) 1–10, https://doi.org/10.1038/s41598-024-63094-9, https://www.nature.com/articles/s41598-024-63094-9.

[29] Z. Zhang, S. Zhang, Y. You, W. Zhu, Diabetic retinopathy grading using hybrid CNN and transformer models on MedMNISTv2 dataset, Computers 14 (5) (2025) 187, https://doi.org/10.3390/computers14050187, https://www.mdpi.com/2073-431X/14/5/187.

[30] W. Liu, Y. Luo, X. Zhang, J. Wang, D. Liang, Y. Zhang, Rotation equivariant convolutions for medical image classification: application to the MedMNISTv2 benchmark, 2025, https://arxiv.org/pdf/2501.09753. arXiv preprint arXiv:2501.09753.

[31] W. Zhang, et al., KAN-integrated MedViT: new state-of-the-art vision transformer for MedMNISTv2 2D classification, 2025, https://arxiv.org/pdf/2502.13693. arXiv preprint arXiv:2502.13693.

[32] S. Saha, A. Das, A. Konar, A.K. Sangaiah, A deep fuzzy rank-based ensemble model for multi-class skin disease classification with DermaMNIST, Sci. Rep. 14 (1) (2024) 1–16, https://doi.org/10.1038/s41598-025-90423-3, https://www.nature.com/articles/s41598-025-90423-3.

[33] Papers With Code, Papers with code: MedMNIST v2 dataset, 2023. https://paperswithcode.com/dataset/medmnist-v2.

[34] Hugging Face, Hugging face: MedMNIST v2 dataset, 2023. https://huggingface.co/datasets/albertvillanova/medmnist-v2.

[35] J. Huang, Y. Peng, L. Hu, A multilayer stacking method base on RFE-SHAP feature selection strategy for recognition of driver's mental load and emotional state, Expert Syst. Appl. 238 (PB) (Mar 2024) 2024), https://doi.org/10.1016/j.eswa.2023.121729

[36] T. Luo, SHAP-based recursive feature elimination and hyperparameter optimization for enhanced financial stock forecasting, in: Proceedings of the 2025 International Conference on Big Data, Artificial Intelligence and Digital Economy, Association for Computing Machinery, New York, NY, USA, 2025, pp. 111–120, https://doi.org/10.1145/3767052.3767070, BDAIE '25.

[37] U. Ahmed, Z. Jiangbin, A. Almogren, M. Sadiq, A.U. Rehman, M.T. Sadiq, J. Choi, Hybrid bagging and boosting with SHAP based feature selection for enhanced predictive modeling in intrusion detection systems, Sci. Rep. 14 (1) (2024) 30532, https://doi.org/10.1038/s41598-024-81151-1

[38] C.E.L. Asry, I. Benchaji, S. Douzi, B.E.L. Ouahidi, Enhancing cybersecurity: a high-performance intrusion detection approach through boosting minority class recognition, PLOS ONE 20 (3) (2025) e0317346.

[39] O.O. Oladimeji, H. Ayaz, I. McLoughlin, S. Unnikrishnan, Mutual information-based radiomic feature selection with SHAP explainability for breast cancer diagnosis, Results Eng. 24 (2024) 103071, https://doi.org/10.1016/j.rineng.2024.103071, https://www.sciencedirect.com/science/article/pii/S2590123024013264.

[40] S. Hong, S. Hong, E. Oh, W.J. Lee, W.K. Jeong, K. Kim, Development of a flexible feature selection framework in radiomics-based prediction modeling: assessment with four real-world datasets, Sci. Rep. 14 (1) (2024) 29297, https://doi.org/10.1038/s41598-024-80863-8

[41] Y. Wang, X. Liu, X. Zhao, Z. Wang, X. Li, D. Sun, A radiomics-based machine learning model and SHAP for predicting spread through air spaces and its prognostic implications in stage I lung adenocarcinoma: a multicenter cohort study, Cancer Imaging 25 (1) (2025) 115, https://doi.org/10.1186/s40644-025-00935-4

[42] J. Jiang, X. Zhang, Z. Yuan, Feature selection for classification with spearman's rank correlation coefficient-based self-information in divergence-based fuzzy rough sets, Expert Syst. Appl. 249 (2024) 123633, https://doi.org/10.1016/j.eswa.2024.123633, https://www.sciencedirect.com/science/article/pii/S0957417424004986.

[43] H. Gong, Y. Li, J. Zhang, B. Zhang, X. Wang, A new filter feature selection algorithm for classification task by ensembling Pearson correlation coefficient and mutual information, Eng. Appl. Artif. Intell. 131 (C) (May 2024), https://doi.org/10.1016/j.engappai.2024.107865

[44] L. Rundo, C. Militello, Image biomarkers and explainable AI: handcrafted features versus deep learned features, European Radiology Experimental 8 (1) (Nov 2024), https://doi.org/10.1186/s41747-024-00529-y

[45] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, vol. 1, 2001, pp. 1, https://doi.org/10.1109/CVPR.2001.990517

[46] L. Putzu, C. Di Ruberto, Rotation invariant co-occurrence matrix features, in: S. Battiato, G. Gallo, R. Schettini, F. Stanco (Eds.), Image Analysis and Processing - ICIAP 2017, Springer International Publishing, Cham, 2017, pp. 391–401.

[47] G. Thibault, B. FERTIL, C. Navarro, S. Pereira, N. Lévy, J. Sequeira, J.-L. Mari, Texture indexes and gray level size zone matrix application to cell nuclei classification, in: 10th International Conference on Pattern Recognition and Information Processing, 2009, pp. 140–145.

[48] N.J. Tustison, J. Gee, Run-length matrices for texture analysis, Insight J. (2011). Article ID: 231. https://doi.org/10.54294/ex0itu

[49] M. Amadasun, R. King, Textural features corresponding to textural properties, IEEE Trans. Syst. Man Cybern. 19 (5) (1989) 1264–1274, https://doi.org/10.1109/21.44046

[50] C. Sun, W.G. Wee, Neighboring gray level dependence matrix for texture classification, Computer Vision, Graphics, and Image Processing 23 (3) (1983) 341–352, https://doi.org/10.1016/0734-189X(83)90032-4, https://www.sciencedirect.com/science/article/pii/0734189X83900324.

[51] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987, https://doi.org/10.1109/TPAMI.2002.1017623

[52] R. Mukundan, S. Ong, P. Lee, Image analysis by Tchebichef moments, IEEE Trans. Image Process. 10 (9) (2001) 1357–1364, https://doi.org/10.1109/83.941859

[53] C.D. Ruberto, L. Putzu, G. Rodriguez, Fast and accurate computation of orthogonal moments for texture analysis, Pattern Recognit. 83 (2018) 498–510, https://

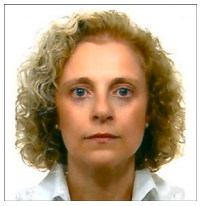
doi.org/10.1016/j.patcog.2018.06.012, https://www.sciencedirect.com/science/article/pii/S003132031830222X.
[54] M.R. Teague, Image analysis via the general theory of moments*, J. Opt. Soc. Am. 70 (8) (1980) 920–930, https://doi.org/10.1364/JOSA.70.000920, https://opg.optica.org/abstract.cfm?URI=josa-70-8-920.
[55] C.-H. Teh, R.T. Chin, On image analysis by the methods of moments, IEEE Trans. Pattern Anal. Mach. Intell. 10 (4) (1988) 496–513.
[56] M. Oujaoura, B. Minaoui, M. Fakir, Image annotation by moments, Moments and Moment Invariants-Theory and Applications 1 (10) (2014) 227–252.
[57] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, 2000, https://doi.org/10.1017/CBO9780511801389
[58] P.-N. Tan, M. Steinbach, A. Karpatne, V. Kumar, Introduction to Data Mining, second ed, Pearson, Upper Saddle River, NJ, 2017.
[59] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, https://doi.org/10.1023/A:1010933404324
[60] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (1) (2006) 3–42, https://doi.org/10.1007/s10994-006-6226-1
[61] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785, KDD '16.
[62] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-Learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[63] S.Y. Park, J.Y. Park, J.W. Park, W.H. Kim, J.Y. Park, H.J. Kim, Unexpected hyperechoic lesions of the breast and their correlations with pathology: a pictorial essay, Ultrasonography 41 (3) (2022) 597–609, https://doi.org/10.14366/usg.21243
[64] A. Athanasiou, A. Tardivon, L. Ollivier, F. Thibault, C. El Khoury, S. Neuenschwander, How to optimize breast ultrasound, Eur. J. Radiol. 69 (1) (2009) 6–13, https://doi.org/10.1016/j.ejrad.2008.07.034
[65] C.M. Sehgal, S.P. Weinstein, P.H. Arger, E.F. Conant, A review of breast ultrasound, Journal of Mammary Gland Biology and Neoplasia 11 (2) (2006) 113–123, https://doi.org/10.1007/s10911-006-9018-0

Author biography

Alessandra Perniciano received her B.Sc. and M.Sc. degrees from the University of Cagliari in 2020 and 2023, respectively. Currently, she is pursuing a Ph.D. at the Department of Mathematics and Computer Science at the University of Cagliari (Italy), focusing on Artificial Intelligence and Machine Learning techniques for analyzing and processing hybrid biomedical data. Her research interests include data mining and machine learning, with a particular emphasis on dimensionality reduction, XAI, class balancing, feature extraction and bioinformatics. She has authored multiple scientific publications and presented her work at international conferences.

Federico Maria Cau obtained his bachelor's and master's degrees from the University of Cagliari, where he also earned a Ph.D. in Mathematics and Computer Science, focusing on the effects of explanation and uncertainty on AI-assisted user decisions. He is currently a postdoctoral researcher at the University of Cagliari. His research interests include AI-assisted decision-making, eXplainable AI, Human-Centered AI, and intelligent interfaces.

Lucio Davide Spano is an Associate Professor at the Department of Mathematics and Computer Science, University of Cagliari, Italy. He received the Ph.D. in Computer Science from his University of Pisa in 2013. He previously worked at ISTI CNR in Pisa within the Human Interfaces in Information Systems Laboratory, contributing to several European FP7 and H2020 projects. He joined the University of Cagliari in 2012, becoming an Associate Professor in 2019. He has been a Visiting Professor at the University of Michigan (USA), University of Toulouse III (France), and University of Primorska (Slovenia). Since July 2024, he has served as Director of the GLab Interdepartmental Research Center, dedicated to the study and promotion of playful practices, pervasive games, and interactive urban experiences. He is currently Vice-Chair of the IFIP TC 13 on Human Computer Interaction, Chair of the IFIP WG 13.4/2.7 on User Interface Engineering and Delegate for Research of the Italian ACM SIGCHI Chapter. His research interests include human–computer interaction, extended reality, end user development, gestural interaction, model based approaches for user interfaces, and intelligent interactive systems. He has served as Steering Committee Chair and General Chair of ACM Engineering Interactive Computing Systems, General Chair of the International Symposium on End User Development, and Program or General Chair for multiple ACM and IFIP conferences. He is Associate Editor of ACM Transactions on Intelligent Interactive Systems, Springer Virtual Reality, and Behavior & Information Technology.



Cecilia Di Ruberto received an M.S. degree cum laude in Computer Science from the University of Salerno and a Ph.D. in Computer Science from the University of Naples. She is an Associate Professor in Computer Science at the Department of Mathematics and Computer Science of the University of Cagliari, Italy. Her research interests include computer vision, image retrieval, medical image analysis, pattern recognition, and machine and deep learning. She has conducted extensive research in microscopic image analysis, particularly in blood smear image analysis for cell counting, malaria parasite detection and classification, and leukemia detection. She is the author of more than 100 scientific papers published in peer-reviewed journals and international conference proceedings.

Andrea Loddo received the B.Sc., M.Sc., and Ph.D. degrees from the University of Cagliari, in 2012, 2014, and 2019, respectively. His Ph.D. thesis addressed blood cell image analysis and classification issues to create new tools for automatic diagnosis to support medical analysis. He is currently an Assistant Professor at the Department of Mathematics and Computer Science, University of Cagliari. He has authored over 50 scientific manuscripts in peer-reviewed journals and international conference proceedings. His research interests include image analysis and processing, computer vision, pattern recognition, and machine and deep learning, with a focus on medical tasks.