Article

Enhancing Medical Image Classification with Unified Model Agnostic Computation and Explainable AI

Elie Neghawi *,† and Yan Liu †

Gina Cody School of Engineering and Computer Science, Concordia University, Montreal, QC H3G 1M8, Canada; yan.liu@concordia.ca

* Correspondence: e_negh@live.concordia.ca

† These authors contributed equally to this work.

Abstract: *Background*: Advances in medical image classification have recently benefited from general augmentation techniques. However, these methods often fall short in performance and interpretability. *Objective:* This paper applies the Unified Model Agnostic Computation (UMAC) framework specifically to the medical domain to demonstrate its utility in this critical area. *Methods*: UMAC is a model-agnostic methodology designed to develop machine learning approaches that integrate seamlessly with various paradigms, including self-supervised, semi-supervised, and supervised learning. By unifying and standardizing computational models and algorithms, UMAC ensures adaptability across different data types and computational environments while incorporating state-of-the-art methodologies. In this study, we integrate UMAC as a plug-and-play module within convolutional neural networks (CNNs) and Transformer architectures, enabling the generation of high-quality representations even with minimal data. *Results*: Our experiments across nine diverse 2D medical image datasets show that UMAC consistently outperforms traditional data augmentation methods, achieving a 1.89% improvement in classification accuracy. *Conclusions*: Additionally, by incorporating explainable AI (XAI) techniques, we enhance model transparency and reliability in decision-making. This study highlights UMAC's potential as a powerful tool for improving both the performance and interpretability of medical image classification models.

Keywords: deep learning; explainable artificial intelligence; convolutional neural networks (CNN); self-supervised machine learning

check for updates

Citation: Neghawi, E.; Liu, Y. Enhancing Medical Image Classification with Unified Model Agnostic Computation and Explainable AI. *AI* **2024**, *5*, 2260–2278. https://doi.org/10.3390/ai5040111

Academic Editors: Affan Yasin, Javed Ali Khan and Lijie Wen

Received: 2 September 2024
Revised: 28 October 2024
Accepted: 28 October 2024
Published: 5 November 2024

1. Introduction

Deep learning methods significantly aid clinicians in rapid examination and accurate diagnosis [1]. However, these methods demand substantial data, which are often scarce in medical contexts. Limited patient data or insufficient medical equipment can lead to biased and overfitting models [1]. Data augmentation, a technique commonly employed to address these issues, is particularly crucial in medical imaging where modalities (e.g., MRI, CT, X-ray) require specialized knowledge, and augmentation can be computationally expensive [2,3]. Traditional image-level augmentations often struggle to introduce sufficient diversity or achieve meaningful semantic transformations. Meanwhile, generative methods, although capable of enhancing diversity, remain complex and computationally intensive [4,5].

Recent advancements in feature-level augmentation techniques provide new approaches for improving the performance of models, especially in medical image classification. Techniques like GuidedMixup [6], which employs saliency maps to preserve the most relevant parts of the image during augmentation, have been shown to reduce label violations by focusing on salient regions. PuzzleMix improves the standard Mixup method by optimizing spatial consistency and preserving local structures, which is crucial in medical imaging tasks involving complex anatomical structures [7]. SaliencyMix further improves upon these techniques by combining saliency-guided augmentations with patch-based mixups to enhance regularization and generalization, even on smaller medical datasets [8].

In addition to these, ResizeMix offers an augmentation strategy that handles the varying resolutions in medical images by resizing patches while mixing them, thus improving performance across different imaging modalities [9]. ReMix addresses class imbalances by synthesizing balanced training data for underrepresented classes, making it particularly valuable in applications like disease classification where certain conditions may be underrepresented [10]. Finally, Co-Mixup introduces a novel approach to jointly mixing images while preserving diversity across multiple sources, making it suitable for small medical datasets [11].

These state-of-the-art augmentation techniques are particularly relevant in medical image classification, where data are often scarce, and diversity in training samples is crucial for model generalization. Building on our previous work [12], where we developed the UMAC framework, we now apply it to the medical domain, integrating these advanced augmentation techniques.

UMAC's model-agnostic approach integrates seamlessly with various learning paradigms, adapting across different data types and environments. It is particularly well-suited to the medical field, where data scarcity, the need for high accuracy, and the demand for interpretability are critical challenges. By generating new models tailored for medical applications and incorporating XAI methodologies, UMAC simplifies the development of effective, reliable, and transparent machine learning solutions. This demonstrates its potential to significantly advance medical research and practice. Given these considerations, our research is guided by the following central question:

Research Question: How can the UMAC framework be effectively applied to the challenging field of medical image classification, addressing issues such as data scarcity, model generalization, and the need for interpretability in clinical settings?

Our main contributions in addressing this question are:

- Developing a high-performance, self-supervised learning module using UMAC techniques, specifically tailored for medical image classification.
- Demonstrating how UMAC techniques provide insights into where performance improvements are achieved.
- Showing that UMAC enhances performance across various dimensions, modalities, and architectures, including CNNs and Transformers, with a particular emphasis on improving data augmentation techniques.

The remainder of this paper is structured as follows: In Section 2, we discuss the challenges in acquiring medical data. Section 3 provides an in-depth discussion of related work, focusing on state-of-the-art techniques in data augmentation and self-supervised learning, particularly in the context of medical imaging. Section 4 introduces the proposed methodology, detailing the integration of UMAC with advanced augmentation strategies. Section 5 describes the datasets and experimental setup used to evaluate the performance of UMAC. The experimental results, comparisons with other cutting-edge methods, and key insights are presented in Section 6. Finally, Section 7 concludes the paper.

2. Challenges in Acquiring Medical Data

Obtaining high-quality, labeled data is a significant challenge in the medical field, especially for training machine learning models. Unlike other domains where data can be easily generated or collected, medical data acquisition is often constrained by several unique factors.

First, data privacy and security is a major concern. Medical data contain sensitive information, and regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in Europe impose strict controls on data collection, storage, and sharing. These regulations are essential to protect patient privacy but can also limit access to large datasets for research purposes [13,14].

Additionally, data labeling and annotation is a complex process in the medical domain. It requires expertise from trained medical professionals, which makes it time-consuming and expensive. The variability in labeling due to subjective interpretations by different experts adds another layer of complexity [15,16].

Another challenge is data diversity and representation. Many medical datasets lack diversity, both in terms of demographic representation and the range of medical conditions covered. Most available datasets are sourced from a limited number of healthcare institutions, often in high-income countries, leading to biases that affect the generalizability of models to broader or underserved populations [17,18].

Furthermore, data size and quality can be a limiting factor. Certain medical conditions are rare, making it difficult to obtain sufficient data to train robust models. The quality of available data may also vary significantly due to differences in imaging equipment, protocols, and patient populations, leading to noisy and inconsistent datasets that hinder model performance and reliability [19,20].

The use of medical data for research raises ethical and legal challenges, particularly around consent and data misuse. Researchers must navigate complex ethical guidelines to ensure data usage aligns with patient rights and expectations, requiring careful collaboration among clinicians, ethicists, and data scientists [21,22].

Finally, data integration and standardization pose significant difficulties. Integrating data from multiple sources, such as electronic health records (EHRs), imaging, and genomic data, requires significant preprocessing, cleaning, and normalization to ensure compatibility and meaningful analysis [23,24].

These challenges highlight the need for innovative approaches like the UMAC framework, which aims to maximize the utility of available data through advanced computational techniques and data augmentation strategies. By addressing the limitations of current datasets, UMAC can improve the performance and generalizability of machine-learning models in the medical field.

3. Related Work

Self-supervised learning (SSL) has emerged as a key approach in machine learning, particularly for leveraging unlabeled data to extract meaningful features. Shurrab et al. [25] provide a comprehensive overview of SSL techniques in medical imaging, discussing their contributions and limitations. While this survey highlights important advancements, it does not delve into the shared foundational components that unify various SSL approaches. Recent studies have demonstrated the effectiveness of SSL in numerous medical imaging tasks. For example, Taleb et al. [26] introduced a multi-task learning framework that integrates SSL with semi-supervised learning, reporting a 7% improvement in classification accuracy for 3D medical images. Similarly, Jamaludin et al. [27] applied SSL to spinal MRI, achieving a 15% increase in feature extraction efficiency, significantly enhancing segmentation accuracy. These findings underscore the potential of SSL to improve performance in situations where labeled data are limited.

In the field of medical image registration, Li and Fan [28] employed self-supervised fully convolutional networks for non-rigid image alignment, reaching an accuracy of 80% without requiring extensive labeled data. Taleb et al. [26] extended SSL to 3D medical imaging, showing a 10% improvement in segmentation tasks, further demonstrating the value of SSL in complex medical imaging scenarios.

XAI techniques have also gained importance alongside SSL, as they provide insights into the decision-making process of machine learning models. LIME [29] and SHAP [30] are widely used XAI methods, though they are often modality-specific. Droste et al. [31] highlighted the need for clinically relevant explainability by developing metrics based on visually salient landmarks. However, these techniques are often limited to specific imaging modalities. The UMAC framework offers a more flexible approach to XAI, enhancing explainability across a range of medical imaging tasks.

Data augmentation plays a critical role in improving model performance, especially in medical imaging, where data are often scarce. GuidedMixup [32], which uses saliency maps to guide interpolation, has demonstrated a 5% improvement in classification accuracy on imbalanced datasets. PuzzleMix [7], which focuses on optimizing spatial consistency during augmentation, improved segmentation accuracy by 8% in tasks involving complex anatomical structures.

Implicit Semantic Data Augmentation (ISDA) [33] enhances model robustness by augmenting features along class-specific semantic directions without generating new samples. ISDA has been shown to improve classification accuracy by 2–4% on datasets such as CIFAR-10 and CIFAR-100 when applied to deep networks like ResNet [33]. Its adaptability to deep learning models makes it highly relevant to medical imaging tasks where labeled data are limited.

Bayesian Random Semantic Data Augmentation (BSDA) [34], introduced in [34], addresses challenges in medical image classification by generating feature augmentations that better represent class distributions. BSDA improved accuracy and area under the ROC curve (AUC) by 2–5% on medical imaging datasets such as BreastMNIST and RetinaMNIST, demonstrating its efficacy in both 2D and 3D modalities [34].

These augmentation methods, while initially developed for general imaging tasks, hold significant potential when applied to medical image classification. The UMAC framework integrates these techniques to improve model generalization and performance across a variety of medical imaging tasks, providing a flexible and powerful tool for the medical research community.

4. Methodology

To apply the UMAC methodology [12] to the medical field, we begin by outlining its essential components and the process involved in developing a UMAC system. The development of a UMAC system follows a structured methodology that efficiently integrates diverse computational models, algorithms, and frameworks. The goal is to create a flexible computation system capable of handling various data types, problems, and computational environments, while also incorporating the latest advancements in methodologies. A key feature of UMAC is the integration of XAI, which enhances transparency and interpretability by providing a plug-and-play structure. This allows developers to clearly understand and modify different components of the model as needed, making it easier to experiment, fine-tune, and improve the system's performance while ensuring that each part of the model is easily interchangeable and its impact is transparent.

When adapting UMAC to the medical field, preprocessing techniques play a critical role in enhancing the performance of self-supervised learning models, particularly in medical image classification. Building on insights from our previous work [12], we developed a systematic approach that unifies and abstracts computational models and algorithms. This ensures adaptability across different data types and computational environments. Specifically, we adopt the Unified Agnostic Computation Process for self-supervised learning, focusing on data augmentation techniques to improve model performance in medical image classification tasks, as shown in Figure 1.

Figure 1 outlines the Unified Agnostic Computation Process for self-supervised learning as applied to medical image classification. The process begins by generating two augmented datasets from the original dataset, which are then processed through two encoders: the Key Encoder and the Queue Encoder. The Key Encoder is represented by parameters θ, while the Queue Encoder is initialized with parameters ξ.

To ensure diversity in the dataset, random parameters σ and σ' are applied during augmentation. After the data pass through the encoders, contrastive loss is calculated by comparing positive pairs (similar images) with negative pairs (dissimilar images). A FIFO queue mechanism is employed to store representations from previous batches, allowing the model to learn from a broader range of examples. This enhances the model's ability to distinguish between different classes.

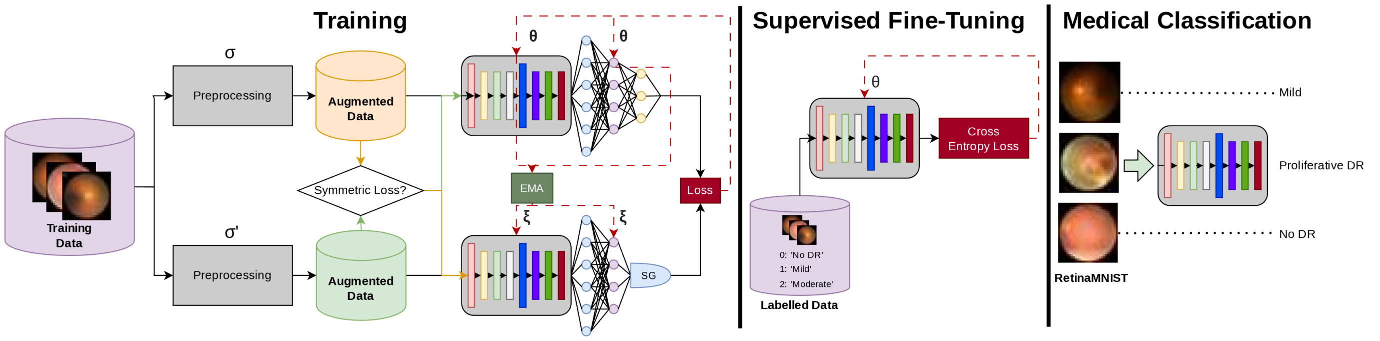

Figure 1. Unified Agnostic Computation Process with self-supervised learning in the medical field. In this context, θ and ξ are parameters, while σ and σ' represent random parameters.

If a symmetric loss function is applied, the augmented input is fed into both network classifiers, which have different parameters: θ for the Key Encoder and ξ for the Queue Encoder. The loss is computed between the two classifiers, typically CNNs or transformers, to capture differences in representation.

Finally, the parameters of the Key Encoder (θ) are used as the starting point for the fine-tuning phase. The output is further fine-tuned using labeled data, which enhances the model's performance in tasks such as medical image classification, as demonstrated in the RetinaMNIST dataset.

4.1. Training

The training phase is a critical component of our system's development, laying the foundation for a robust computational model. This stage is meticulously designed to enhance model performance through various preprocessing and network classification strategies.

- **Preprocessing:** The detailed procedure for image data augmentation is outlined in Algorithm 1. This algorithm aims to increase the dataset size while preserving the original label distribution. Given an original set of images $X = \{x^1, x^2, x^3, \ldots, x^n\}$ and corresponding labels $Y = \{y^1, y^2, y^3, \ldots, y^n\}$, where each label $y^i \in \{0, 1, 2, \ldots, u\}$, the objective is to expand the dataset by a factor α.
  To achieve this, the algorithm performs the following steps:
  - Initialization: Calculate the desired size of the augmented datasets:

    $$m = \lceil n \times \alpha \rceil$$

    Initialize empty sets S_1 and S_2 for storing augmented images. Compute the label distribution $P(y)$ from Y to maintain the original distribution in the augmented datasets.
  - Augmentation Strategy: For each image in the original dataset X, apply a series of random augmentations to create new images. The augmentation function A is defined using parameters generated randomly:

    $$X_{\text{aug}} = A(x)$$

    Each image is augmented at least once to ensure diversity in the augmented datasets.

    $$\text{If } C(y) < p_y \times m, \text{ then augment image } x.$$
  - Final Datasets: The resulting augmented datasets S_1 and S_2 are expanded to the target size m, with label distributions that closely follow the original dataset's

distribution. This process enhances the model's ability to generalize and handle diverse data scenarios.

To further illustrate this augmentation process, Figure 2 provides a concrete example using a DermaMNIST image. This example demonstrates the series of transformations applied during augmentation, such as random color shifts and spatial adjustments, highlighting how these changes ensure diversity in the training data.

Algorithm 1 Image Data Augmentation with Size Increase

```
Require:
    | Original set of images $X = \{x^1, x^2, x^3, \dots, x^n\}$
    | Original set of labels $Y = \{y^1, y^2, y^3, \dots, y^n\}$, where $y^i \in \{0, 1, 2, \dots, u\}$
    | Desired augmentation factor $\alpha$
Ensure: Augmented image sets $S_1, S_2$ with size increased by factor $\alpha$
$S_1, S_2 \leftarrow \emptyset$
$m \leftarrow \lceil n \times \alpha \rceil$
$P(y) = \{p_0, p_1, \dots, p_u\}$ from $Y$
$C(y) = \{c_0 = 0, c_1 = 0, \dots, c_u = 0\}$
for $i = 1$ to 2 do
    for $j = 1$ to $n$ do
        $\sigma \leftarrow \text{rand}()$
        $B_i, N_i, C_i, O_i \leftarrow \sigma_i$
        $c_{ik} \leftarrow \text{rand}(), \forall k \in [1, k_{\text{color}}]$
        $o_{ik} \leftarrow \text{rand}(), \forall k \in [1, k_{\text{spatial}}]$
        $A_i \leftarrow \text{createAugmentationFunction}(B_i, N_i, C_i, O_i, c_{i1}, \dots, c_{ik}, o_{i1}, \dots, o_{ik})$
        $X_{\text{aug}} \leftarrow A(x^j)$
        $S_i.\text{add}(X_{\text{aug}})$
    end for
    while size of $S_i < m$ do
        Select a random number $r \in \{1, 2, \dots, n\}$
        Determine label $y_r \leftarrow y^r$
        if $C(y_r) < p_{y_r} \times m$ then
            Obtain image $x^r$ from $X$
            $\sigma \leftarrow \text{rand}()$
            $B_i, N_i, C_i, O_i \leftarrow \sigma_i$
            $c_{ik} \leftarrow \text{rand}(), \forall k \in [1, k_{\text{color}}]$
            $o_{ik} \leftarrow \text{rand}(), \forall k \in [1, k_{\text{spatial}}]$
            $A_i \leftarrow \text{createAugmentationFunction}(B_i, N_i, C_i, O_i, c_{i1}, \dots, c_{ik}, o_{i1}, \dots, o_{ik})$
            $X_{\text{aug}} \leftarrow A(x^r)$
            $S_i.\text{add}(X_{\text{aug}})$
        end if
    end while
end for
return $S_1, S_2$
```

- **Network Classifier:** The prevailing models predominantly incorporate two encoders, punctuated with key components:
  - **Encoder Architecture**: The encoder can be any of the popular CNN architectures such as ResNet [35], ResNeXt [36], or DenseNet [37]. Larger architectures generally yield superior results, albeit at heightened computational costs. Specific components, especially the queuing of representations, can judiciously curtail the model's size and batch requisites without compromising performance. The minibatch size is set to 128 by default for each of these CNN architectures but can be adjusted if needed. However, we did not find significant differences when altering the minibatch size, which is a limitation of our experiment. In addition to CNN-based encoders, Transformer-based architectures like Vision Transformer

(ViT) [38] can also be utilized for image classification tasks, providing a versatile option for the encoders based on self-attention mechanisms.

- **Auxiliary Components**:
    - **MLPs**: Post-encoder MLPs are non-negotiable. A deeper MLP on the Key Encoder relative to the Queue Encoder is quintessential. A more intricate and expansive Encoder Architecture mandates a correspondingly profound MLP.
    - **Representation Queue**: The representation queue, while adhering to the FIFO principle, is also influenced by the learning rate of the key encoder. A higher learning rate necessitates a smaller queue due to rapid weight updates rendering stored representations obsolete swiftly. Conversely, with a slower learning rate, representations evolve more gradually, permitting a more extensive queue. Mathematically, the FIFO operation in terms of batches, influenced by a hyperparameter h, can be articulated as follows:

      $$Q_{b+1} = \text{Append}(Q_b, k_{b+1}) - \text{Remove}(Q_b, h)$$

      where Q_b symbolizes the queue's state at batch b, k_{b+1} denotes the key representations of the newly processed batch, and h indicates the number of batch sizes' worth of representations to be removed. Adjusting h allows for fine-tuning the refresh rate of the queue, providing a balance between queue longevity and representational freshness.
    - **Exponential Moving Average (EMA)**: A straightforward procedure where solely the Key Encoder's value undergoes modifications, employing EMA to contemporaneously update the Queue Encoder's parameters.

Let θ be the randomly initialized parameters of the Key Encoder and ξ be the parameters of the Queue Encoder. The training process involves the two augmented datasets, and a backward propagation updates θ while EMA is used to update ξ:

$$\xi \leftarrow \beta\xi + (1-\beta)\theta$$

where β is the EMA decay rate. This process is illustrated in Figure 1.

- **Loss Function:** The nature of the loss function plays a pivotal role in dictating the interaction between augmented data and the encoders, determining if both augmented datasets traverse both encoders or just one. Coupled with this, the role of queuing becomes evident:
    - **Contrastive Loss and Queuing:** In architectures that employ representation queuing, the contrastive loss is especially effective. A queue that captures representations from previous batches enables the network not only to contrast against the positive pair but also against a vast array of negatives. This extensive negative sampling sharpens the encoder's ability to discern between semantically close and diverse data points. In the absence of such a queue, the contrastive loss mainly depends on positive pairs, potentially overlooking the fine nuances provided by many negative samples. As such, leveraging the contrastive loss alongside a queue not only expands the range of representations but also enriches the learning process, setting a more comprehensive contrastive context.
    - **Non-contrastive Loss and Queuing:** For architectures employing a non-contrastive loss, there is a tendency to sidestep processing both augmented datasets through the two (or 'twin') encoders, choosing a more linear path. While this simplifies the computational trajectory, it might forgo the advantages of contrasting augmented views in a dense representational setting.

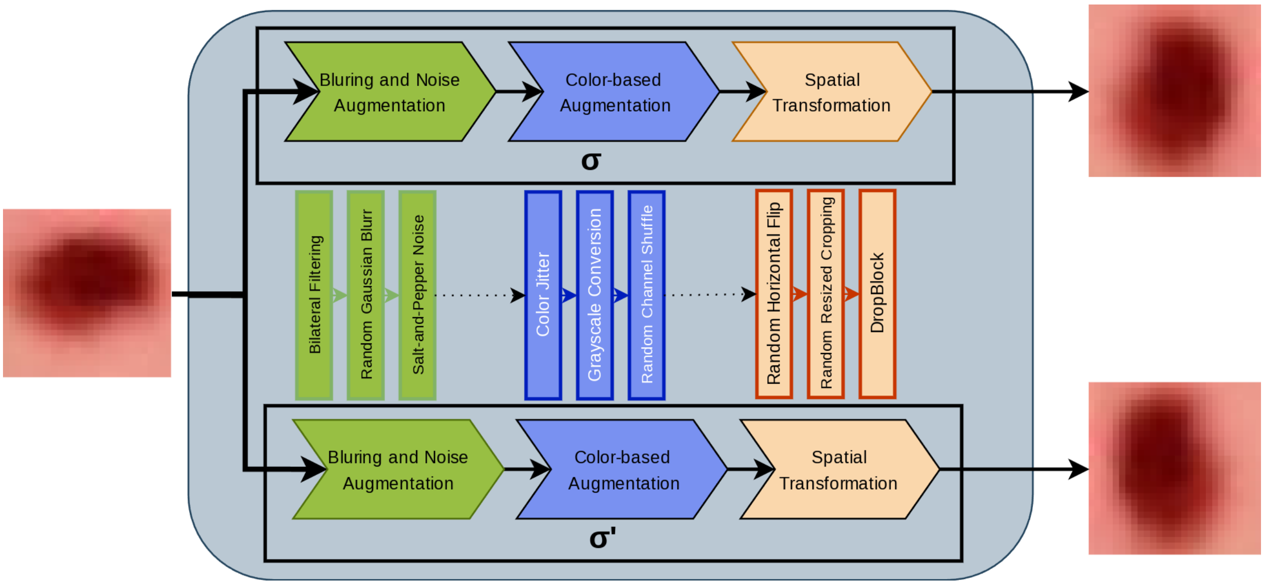

Figure 2. Example of Augmentation Function applied to a DermaMNIST image, showcasing color shifts and spatial transformations.

4.2. Supervised Fine-Tuning

Before we start the fine-tuning process, it is important to note that θ is not randomly initialized for this phase. Instead, we leverage the parameters from the training mode in the previous step. This allows us to benefit from the previously learned representations, enhancing the supervised fine-tuning process.

The process of supervised fine-tuning fundamentally revolves around equipping the key encoder with capabilities to handle labeled data. At the heart of this process lies the widely adopted cross-entropy loss, which serves as the objective function for this phase of training.

Essentially, this entails a basic supervised training regimen for the key encoder. In contrast to the unsupervised or self-supervised paradigms previously discussed, here, the model explicitly learns from data that carries associated labels. Notably, only a percentage of the data, which is labeled, is employed for this fine-tuning. Often, this subset of labeled data is particularly used for benchmarking purposes to assess and compare model performances.

To facilitate the training, a softmax layer is appended at the tail end of the encoder. This layer's primary function is to produce probability distributions over the possible classes for each input sample.

Mathematically, if C denotes the number of classes, the output of the key encoder is fed into a softmax function adjusted to yield a C-dimensional vector. This vector essentially captures the likelihood of the input sample belonging to each of the C classes. The formula can be expressed as follows:

$$\text{Softmax}(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}}$$

where x is the output of the key encoder and i ranges from 1 to C.

The cross-entropy loss, often used in classification tasks, measures the difference between the true labels and the predicted probability distributions. For a single sample, the cross-entropy loss $H(y, \hat{y})$ between the true label y and the predicted probability distribution $\hat{y}$ is given by the following:

$$H(y, \hat{y}) = -\sum_{c=1}^{C} y_c \log(\hat{y}_c)$$

where C is the number of classes, y_c is the true label for class c (often a binary indicator whether the sample belongs to class c or not), and $\hat{y}_c$ is the predicted probability for class c.

The produced probabilities are then contrasted with the true labels using the cross-entropy loss to guide the fine-tuning of the encoder. The loss is then backpropagated through the encoder to update its parameters.

Upon successful fine-tuning using the labeled data subset, the trained key encoder is subsequently applied to medical image classification tasks, as demonstrated in Figure 1 for the RetinaMNIST dataset [39].

5. Experimental Setup

The primary objective of this study is to evaluate the effectiveness of the UMAC framework in medical image classification tasks using the MedMNIST+ dataset. Specifically, we aim to assess UMAC's performance across different data modalities, neural network architectures, and augmentation strategies. Our goal is to determine whether UMAC can enhance model performance, improve generalization, and maintain interpretability compared to state-of-the-art methods.

The observation targets include key performance metrics such as accuracy (ACC) and the area under the ROC curve (AUC) across multiple datasets with varying complexities and data modalities (e.g., X-ray, OCT, Ultrasound). By focusing on these metrics, we observe the effectiveness of UMAC in handling diverse classification tasks, including binary classification, multi-class classification, ordinal regression, and multi-label classification.

When presenting the results, we aim to address several critical aspects of our research questions. We investigate whether the UMAC framework offers superior performance compared to existing state-of-the-art methods across different medical image datasets. Additionally, we evaluate UMAC's adaptability to various neural network architectures and its impact on training stability and model robustness. Furthermore, we explore the effectiveness of different data augmentation strategies within UMAC, particularly their role in enhancing model generalization to unseen data.

In this section, we empirically validate the proposed algorithm using MedMNIST+ [39], a large-scale collection of standardized biomedical images. Our evaluation strategy covers several crucial aspects: comparison with state-of-the-art methods, effectiveness across different modalities and dimensions, and adaptability to various neural network architectures. Additionally, we conducted ablation experiments, hyperparameter analysis, and visualizations of deep features.

5.1. Datasets

The MedMNIST+ [39] dataset comprises twelve pre-processed 2D datasets and six pre-processed 3D datasets from selected sources covering primary data modalities (e.g., X-ray, OCT, Ultrasound, CT, Electron Microscope), diverse classification tasks (binary/multi-class, ordinal regression, and multi-label), and data scales (from 100 to 100,000) [39]. We selected five 2D medical image datasets in MedMNIST+ [39] covering different modalities. For more details on the dataset, please refer to Table 1. We selected these 2D images due to computational restrictions, which we will discuss in the next subsection.

Table 1. Summary of Selected 2D Medical Image Datasets. The columns represent the number of samples for Training (T), Validation (V), and Test (Te), and the number of classes (C).

Dataset	**Data Modality**	**Tasks (C)**	**Samples (T/V/Te)**
BreastMNIST	Ultrasound	Binary Classification (2)	546/78/156
DermaMNIST	Dermatology	Multi-class Classification (7)	7000/1500/2000
RetinaMNIST	OCT	Multi-class Classification (5)	1000/200/400
ChestMNIST	X-ray	Multi-label Classification (14)	78,468/11,219/22,435
PneumoniaMNIST	X-ray	Binary Classification (2)	4708/524/624

Each dataset has a distinct class distribution, as detailed in Table 2. The table provides an overview of the exact number of samples in each class and their corresponding percentage of the total dataset. Understanding this class distribution is crucial for evaluating potential biases and imbalances that may impact the performance of machine learning models.

Table 2. Detailed Class Distribution for Selected 2D Medical Image Datasets. The table includes the number of samples in each class and the corresponding percentage of total samples for each dataset.

Dataset	Class	Number of Samples	Percentage (%)
BreastMNIST	Benign	348	63.74%
	Malignant	198	36.26%
DermaMNIST	Melanocytic nevi	6705	67.05%
	Melanoma	111	1.11%
	Benign keratosis	514	5.14%
	Basal cell carcinoma	327	3.27%
	Actinic keratoses	239	2.39%
	Vascular lesions	142	1.42%
	Dermatofibroma	62	0.62%
RetinaMNIST	No DR	535	53.50%
	Mild NPDR	153	15.30%
	Moderate NPDR	158	15.80%
	Severe NPDR	83	8.30%
	Proliferative DR	71	7.10%
ChestMNIST	Atelectasis	13,078	16.67%
	Cardiomegaly	2662	3.39%
	Effusion	10,335	13.17%
	Infiltration	1087	1.39%
	Mass	1891	2.41%
	Nodule	2051	2.61%
	Pneumonia	984	1.25%
	Pneumothorax	2926	3.73%
	Consolidation	1221	1.56%
	Edema	2531	3.22%
	Emphysema	1704	2.17%
	Fibrosis	855	1.09%
	Pleural Thickening	1515	1.93%
	Hernia	164	0.21%
PneumoniaMNIST	Non-pneumonia	3875	82.34%
	Pneumonia	1333	17.66%

To provide a visual overview of the selected datasets, Figure 3 presents a set of sample images from BreastMNIST, DermaMNIST, RetinaMNIST, ChestMNIST, and PneumoniaMNIST. These images exemplify the diversity of modalities, including X-ray, OCT, Ultrasound, and Dermatology. Visualizing these images is essential for understanding the unique characteristics of each dataset, including image resolution and variability in appearance between classes.

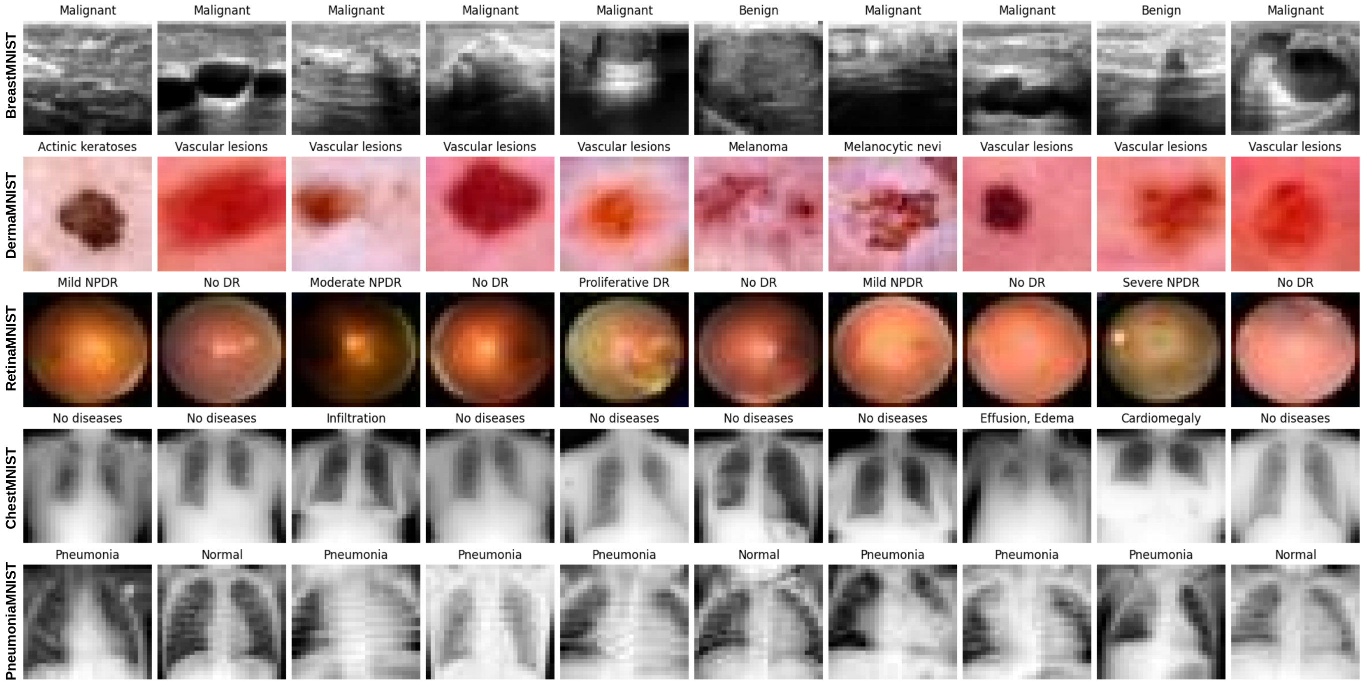

Figure 3. Sample images from the MedMNIST datasets, including examples from BreastMNIST, DermaMNIST, RetinaMNIST, ChestMNIST, and PneumoniaMNIST.

5.2. Implementation Details and Evaluation Protocols

UMAC was applied to the MedMNIST+ dataset to determine hyperparameters, with results reported based on the test set. Often, the randomness in model selection is overlooked, leading to misleading conclusions about method performance [40]. To ensure robustness, the process was repeated three times with different random seeds. Each reported metric is the average of these repetitions, along with their estimated standard error. The evaluation metrics used are the area under the AUC and ACC.

UMAC was implemented using PyTorch Stable (2.3.1) and Torchvision 0.18.1, with experiments conducted on a single NVIDIA RTX 4090 GPU and an Intel 13900k CPU. This implementation is based on the BYOL architecture. The 2D images used were sized at 224 × 224 pixels. Consistent training configurations were maintained across all experiments to ensure fairness. The AdamW optimizer [41] was employed with a learning rate of 0.001, and a learning rate warmup strategy was applied during the initial five epochs of training.

6. Results

We conducted an evaluation of cutting-edge methods using five 2D medical image datasets from MedMNIST2D [39], which include BreastMNIST, DermaMNIST, RetinaMNIST, ChestMNIST, and PneumoniaMNIST, totaling 130,858 samples. Our comparison included UMAC with the BYOL design against leading augmentation techniques such as BSDA [34], ISDA [33], Cutout [42], Mixup [43], and CutMix [44] across these datasets. UMAC, through its preset tasks, serves as an augmentation technique, which is used to train the model and update the parameters more effectively than starting from random initialization, especially given the extensive amount of training data available. This method aligns with the principles of self-supervised learning, where models are pre-trained on specific tasks to enhance performance on the main dataset.

6.1. ACC Results

Table 3 demonstrates that UMAC with Self-Supervised learning is the top-performing method, achieving the highest average accuracy of 81.97% across all evaluated datasets. Although ISDA and BSDA also show strong performance with an average accuracy of 79.58% and 80.08%, respectively, they are slightly less consistent compared to UMAC with Self-Supervised learning. This highlights the benefits of pretrained parameter updates for medical images and underscores the superior performance of UMAC with Self-Supervised learning over ISDA and BSDA. For instance, UMAC with Self-Supervised learning achieved 96.49% accuracy on ChestMNIST and 88.86% on BreastMNIST, whereas BSDA achieved 95.78% and 86.1% on these datasets, respectively. While methods like CutMix [42], CutOut [42], and MixUp [43] provide comparable results with average accuracies of 77.66%, 78.80%, and 76.53%, respectively, none consistently surpass the performance of ISDA, BSDA, and UMAC with Self-Supervised learning. RetinaMNIST remains the most challenging dataset, with all methods exhibiting lower accuracy levels around 50–53%, such as ISDA at 52.6% and UMAC with Self-Supervised learning at 51.3%. BSDA leads in this dataset with 53.3%, though UMAC with Self-Supervised learning will be improved in the next subsection.

Table 3. ACC Performance Comparison of Selected Methods on the Five Different MedMNIST2D Datasets. The "Official" method refers to the baseline provided by MedMNIST+ [39]. Bold values indicate the highest performance, while underlined values indicate the runner-up.

Method	Breast	Derma	Retina	Pneumonia	Chest	Avg
Official	83.3	75.4	49.3	86.4	94.4	77.76
Mixup	83.5 ± 3.2	76.6 ± 0.9	51.3 ± 0.9	81.6 ± 6.1	89.63 ± 3.1	76.53
Cutout	86.3 ± 3.7	75.6 ± 0.1	51.5 ± 4.9	86.1 ± 0.5	94.5 ± 2.8	78.80
CutMix	84.6 ± 0.6	76.3 ± 0.5	52.2 ± 1.5	83.6 ± 7.5	91.6 ± 2.1	77.66
ISDA	86.1 ± 1.0	76.7 ± 0.4	52.6 ± 1.5	87.2 ± 3.7	95.3 ± 2.3	79.58
BSDA	86.1 ± 1.5	76.4 ± 0.8	**53.3 ± 0.1**	88.8 ± 1.2	95.78 ± 2.1	80.08
UMAC (Ours)	**88.86 ± 2.3**	**78.89 ± 0.72**	51.3 ± 1.1	**90.3 ± 2.3**	**96.49 ± 2.7**	**81.97**

6.2. AUC Results

The AUC provides a measure of a model's ability to distinguish between classes. A higher AUC indicates better performance, with a value of 1 representing a perfect classifier. In scenarios with class imbalances, AUC is a more reliable metric than simple accuracy because it accounts for the true positive rate (sensitivity) and false positive rate (1 − specificity). Table 4 presents the AUC performance comparison across five MedMNIST2D datasets.

As shown, UMAC achieved the highest average AUC of 90.72, excelling particularly on the ChestMNIST dataset with a score of 96.8. BSDA also performed well with an average AUC of 90.60, securing the second-highest scores across most datasets. Although Mixup and ISDA provided competitive results, they did not match the consistently high performance of UMAC and BSDA.

Table 4. AUC Performance Comparison of Selected Methods on the Five Different MedMNIST2D Datasets. The "Official" method refers to the baseline provided by MedMNIST+ [39] . Bold values indicate the highest performance, while underlined values indicate the runner-up.

Method	Breast	Derma	Retina	Pneumonia	Chest	Avg
Official	89.1	92.0	71.0	95.6	94.4	88.42
Mixup	89.5 ± 1.2	92.7 ± 0.5	71.9 ± 1.3	95.8 ± 0.4	89.63 ± 3.1	87.51
Cutout	91.1 ± 1.5	93.0 ± 0.5	72.5 ± 1.4	95.9 ± 0.6	94.5 ± 2.8	89.40
CutMix	90.7 ± 1.0	92.9 ± 0.4	73.4 ± 1.3	96.4 ± 0.6	91.6 ± 2.1	89.80
ISDA	89.3 ± 2.0	93.0 ± 0.4	74.1 ± 1.4	95.0 ± 1.1	95.3 ± 2.3	89.34
BSDA	91.4 ± 0.2	93.1 ± 0.2	**75.0 ± 0.7**	95.7 ± 0.2	95.78 ± 2.1	90.60
UMAC (Ours)	**93.8 ± 2.1**	**93.2 ± 0.5**	73.2 ± 1.1	**96.6 ± 2.0**	**96.8 ± 1.5**	**90.72**

6.3. F1-Score Results

The F1-score results provide a more balanced assessment of model performance, especially in scenarios with class imbalances. Table 5 shows that UMAC with Self-Supervised learning achieved the highest average F1-score of 82.54%, demonstrating its superior ability to handle both precision and recall. While BSDA and ISDA also performed well with average F1-scores of 80.62% and 79.85%, respectively, they fell short compared to UMAC's consistent performance across datasets. The highest F1-score was achieved by UMAC on ChestMNIST (96.81%).

Table 5. F1-Score Performance Comparison of Selected Methods on the Five Different MedMNIST2D Datasets. The best results are bold-faced.

Method	Breast	Derma	Retina	Pneumonia	Chest	Avg
Mixup	82.6	75.8	51.0	80.5	89.1	75.8
Cutout	85.5	76.3	51.9	85.9	93.6	78.6
CutMix	84.0	75.9	51.7	82.9	91.4	77.2
ISDA	85.7	77.2	52.3	87.1	95.1	79.85
BSDA	86.0	77.5	**53.2**	88.6	95.4	80.62
UMAC (Ours)	**88.4**	**78.6**	52.0	**90.1**	**96.81**	**82.54**

6.4. Evaluation of Different Network Classifiers with UMAC

In this section, we evaluate the performance of various convolutional neural networks and vision transformer architectures when using the UMAC framework on the PneumoniaMNIST dataset. Table 6 presents the results of applying UMAC and BSDA to several widely used models, including ResNet, DenseNet, and ViT, alongside the baseline performance without augmentation. The results demonstrate that UMAC consistently improves upon both the baseline and BSDA across most networks, in terms of both ACC and AUC.

UMAC shows notable improvements over BSDA and baseline in almost all network architectures. For example, in ResNet-18, UMAC increases the accuracy by 9.8% and AUC by 2.0% compared to the baseline, and by 3.1% and 1.4% compared to BSDA. In addition to accuracy and AUC, we also measure the additional computational overhead introduced by UMAC. Although UMAC increases the training time marginally compared to BSDA, the performance gains justify the added complexity, especially in high-stakes domains such as medical image classification.

Table 6. Evaluation of Baseline, BSDA, and UMAC on different convolutional neural networks using the test set of PneumoniaMNIST. The best results are bold-faced, and the number in brackets denotes the performance improvements achieved by UMAC over BSDA. The last column shows the additional time (AT) introduced by BSDA and UMAC.

Network	ACC (%)			AUC (%)			AT (%)	
	Baseline	BSDA	UMAC	Baseline	BSDA	UMAC	BSDA	UMAC
ResNet-18	82.1	88.8	**91.9** (+3.1)	95.1	95.7	**97.1** (+1.4)	3.7	4.5
ResNet-50	87.0	86.3	**88.7** (+2.4)	96.8	96.9	**97.3** (+0.4)	5.9	6.7
DenseNet-121	84.9	89.4	**91.1** (+1.7)	96.6	96.9	**97.5** (+0.6)	1.5	2.1
ViT-T	82.9	86.0	**87.9** (+1.9)	94.9	96.0	**97.2** (+1.2)	7.5	8.4
ViT-S	81.1	87.2	**89.0** (+1.8)	95.3	95.9	**97.0** (+1.1)	5.8	6.3
ViT-B	81.8	86.8	**88.3** (+1.5)	94.1	95.2	**96.3** (+1.1)	2.3	3.0
Swin-T	73.6	77.0	**79.3** (+2.3)	87.3	92.0	**93.8** (+1.8)	1.4	2.0
Swin-S	63.9	71.7	**74.1** (+2.4)	81.9	90.6	**92.4** (+1.8)	2.1	3.0
Swin-B	62.5	62.5	**65.2** (+2.7)	88.3	88.3	**89.9** (+1.6)	1.3	2.2

As shown in Table 6, UMAC offers consistent improvements over both the baseline and BSDA. The largest gain in accuracy (9.8%) was observed with ResNet-18, demonstrating UMAC's capability to boost performance across different network architectures. In addition, Vision Transformers (ViT and Swin) also benefitted from UMAC, with notable improvements in both accuracy and AUC.

Despite the slight increase in training time due to the added complexity of UMAC, the significant performance improvements make it a valuable enhancement, particularly in scenarios where model accuracy and reliability are of utmost importance, such as medical diagnosis.

6.5. Comparison Experiments with the Use of Multiple Datasets for Training

Building on our analysis of ACC and AUC, we explored the impact of using multiple MedMNIST2D datasets for pretraining, as illustrated in Figure 4. Our results, detailed in Table 7, demonstrate that using multiple datasets for training (UMAC-MD) yields improvements over training with only one dataset (UMAC-1D). This improvement leverages the pre-set tasks used in training the θ parameters, where the augmentation of images is compared against these tasks.

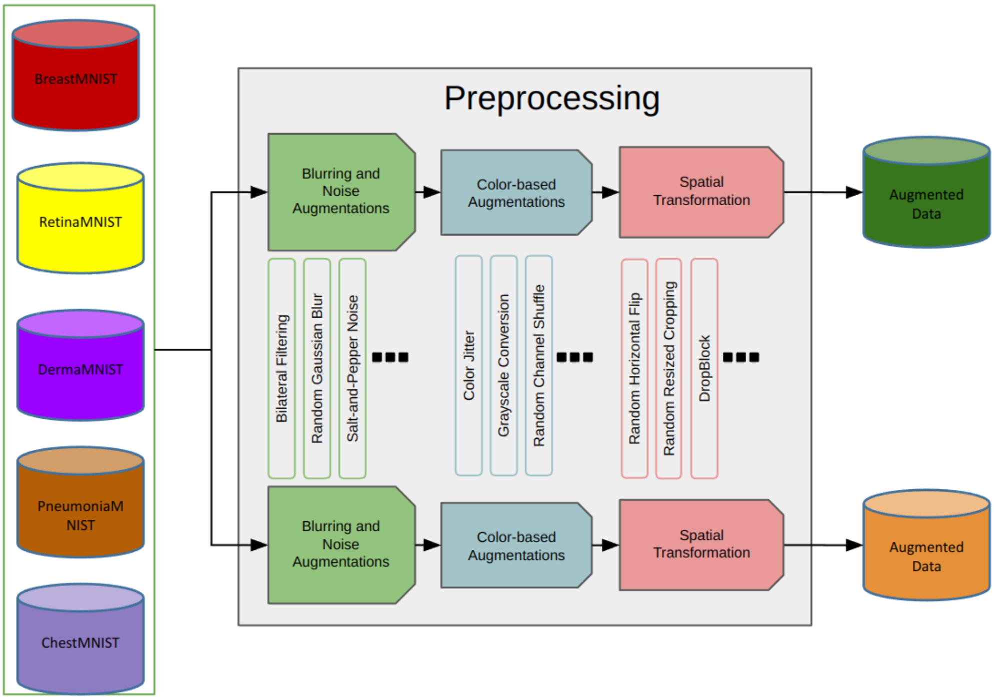

Figure 4. UMAC training with Multiple MedMNIST2D Datasets.

UMAC-1D Training Details: UMAC-1D was trained and tested on a single dataset. For instance, when evaluating BreastMNIST, the model was trained solely on BreastMNIST and tested on the same dataset. The learning rate remained consistent throughout the training and testing process. This approach followed a standard supervised learning setup on one dataset, without leveraging data from other datasets.

UMAC-MD Training Details: In contrast, UMAC-MD leverages pretraining on multiple MedMNIST2D datasets. During pretraining, the model is trained on auxiliary datasets (e.g., DermaMNIST, RetinaMNIST, ChestMNIST, and PneumoniaMNIST) using a lower learning rate, typically reduced by a factor (e.g., 0.1). This reduced learning rate allows the model to learn from the auxiliary datasets without overfitting any specific one.

After pretraining, the model switches to the target dataset (e.g., BreastMNIST) for the main training phase. During this training on the target dataset, the learning rate is increased back to the standard value, allowing the model to focus more on optimizing for the target data. Fine-tuning is also applied during this phase to further refine the model based on the specific features of the target dataset. The combination of pretraining on multiple datasets with a lower learning rate and fine-tuning on the target dataset helps the model generalize better and achieve superior performance.

The results in Table 7 show that UMAC-MD, with its multi-dataset pretraining strategy, yields the highest average accuracy of 82.85%.

UMAC-MD's approach of using multiple datasets for pretraining, followed by targeted training and fine-tuning with an increased learning rate on the dataset of interest, offers significant advantages over both BSDA and UMAC-1D. This method allows the model to learn from a variety of data sources while still optimizing for a specific dataset during the final training and fine-tuning phases.

Table 7. ACC Performance Comparison of Selected Methods on the Five Different MedMNIST2D Datasets, including UMAC with One or More Datasets. The highest accuracy is bold-faced, while the second-highest (runner-up) is underlined.

Method	Breast	Derma	Retina	Pneumonia	Chest	Avg
Official	83.3	75.4	49.3	86.4	94.4	77.76
ISDA	86.1 ± 1.0	76.7 ± 0.4	52.6 ± 1.5	87.2 ± 3.7	95.3 ± 2.3	79.58
BSDA	86.1 ± 1.5	76.4 ± 0.8	53.3 ± 0.1	88.8 ± 1.2	95.78 ± 2.1	80.08
UMAC-1D	88.86 ± 2.3	78.89 ± 0.72	51.3 ± 1.1	90.3 ± 2.3	96.49 ± 2.7	81.17
UMAC-MD	**90.13 ± 1.2**	**80.03 ± 0.87**	**54.7 ± 0.7**	**93.0 ± 1.7**	**97.18 ± 1.7**	**82.85**

6.6. Comparing the Augmentation Factor α

As shown in Table 8, the average augmentation factor α for UMAC methods significantly reduced when datasets were combined for training. Specifically, when we incorporated 1000 images from each of the remaining datasets into the training process, we observed a noticeable decrease in the required augmentation factor.

Table 8. Best Augmentation Factors α for Selected Methods on the Five Different MedMNIST2D Datasets.

Method	Breast	Derma	Retina	Pneumonia	Chest	Avg
UMAC-1D	7.4	3.7	5.8	1.3	2.3	4.10
UMAC-MD	6.5	3.1	5.3	1.7	2.1	3.74

This suggests that integrating diverse datasets can enhance the robustness of the model, thereby reducing the need for extensive data augmentation to achieve optimal performance.

6.7. Summary and Implications

The experimental results demonstrate how the UMAC framework effectively addresses the challenges outlined in Section 2 and provides answers to the central research question posed in this study.

First, the UMAC framework helps overcome the challenge of data scarcity in medical imaging by employing self-supervised learning techniques and feature-level data augmentation. By pre-training models on multiple datasets (as shown in Figure 4) and utilizing pre-set tasks, UMAC reduces the reliance on large labeled datasets, thus mitigating the difficulty of acquiring annotated medical data. This approach allows the model to learn robust representations even with limited data, directly addressing the issue of data scarcity and enhancing model generalization.

Second, UMAC enhances the diversity and quality of training data through advanced data augmentation strategies. The reduction in the augmentation factor α (Table 8) when combining datasets indicates that UMAC can effectively leverage diverse data sources to improve model robustness without the need for extensive, manually-tuned data augmentation. This capability addresses the challenge of limited data diversity and improves the model's ability to generalize to a broader range of medical conditions.

Third, by integrating feature-level augmentation methods that focus on both semantic direction and strength, UMAC maintains high performance across different modalities and neural network architectures, including CNN. This adaptability is crucial in the medical field, where different imaging modalities require specialized handling to ensure accurate diagnosis. The results shown in Tables 3–5 highlight UMAC's consistent outperformance across various datasets and metrics, confirming its effectiveness in enhancing model performance and interpretability.

The F1-score results further demonstrate UMAC's ability to handle both precision and recall effectively. As shown in Table 5, UMAC achieved the highest average F1-score of 82.54%, surpassing alternative methods such as BSDA (80.62%) and ISDA (79.85%). Particularly notable is UMAC's performance on the ChestMNIST dataset, where it reached an F1-score of 96.81%, indicating its superior capability in addressing class imbalance and achieving high performance in both precision and recall. This performance underscores UMAC's potential for real-world medical applications where both false positives and false negatives must be minimized.

Finally, UMAC provides a structured approach to machine learning operations, offering a clear computation graph that outlines the flow of data and processing steps. This structure helps to understand how different components and algorithms interact within the model, ensuring that the machine learning process is well-organized and consistent. While UMAC does not directly enhance transparency in terms of model decision-making, it does offer a well-defined framework that aids in understanding the overall operation of the model. This structured approach aligns with the research question's focus on improving the reliability and trustworthiness of machine learning models in medical contexts by clarifying the computational processes involved.

6.8. Limitations of the Study

While the results presented in this study demonstrate the potential of UMAC in enhancing model performance on 2D medical image datasets, there are several limitations that should be acknowledged.

First, this study is limited to 2D medical images, such as those found in the MedMNIST2D collection. Although UMAC has shown significant improvements in accuracy and AUC within this domain, we have not extended our experiments to 3D medical images, which are prevalent in many real-world applications such as MRI and CT scans. The absence of evaluation on 3D datasets leaves the question open as to how well UMAC would perform in more complex, three-dimensional modalities where spatial relationships across different planes are crucial. Future work should investigate the applicability of UMAC to 3D medical images to determine its generalizability beyond 2D imaging tasks.

Second, while UMAC demonstrated impressive performance across various datasets, the computational complexity of the framework could be a barrier to its adoption in resource-constrained environments. The integration of multiple datasets and advanced feature-level augmentations requires significant computational power, which may not be available in all healthcare settings, particularly in low-resource environments. This limitation highlights the need for optimization strategies that can reduce computational overhead without compromising performance.

By addressing these limitations, future work can further validate and expand the potential of UMAC, ensuring that it remains a robust and flexible tool for a wide range of medical imaging tasks.

7. Conclusions

In this study, we have demonstrated the effectiveness of the UMAC method in achieving superior performance across various 2D medical image datasets from MedMNIST2D. By employing UMAC with self-supervised learning, we have shown significant improvements in both accuracy and AUC compared to traditional augmentation techniques such as BSDA, ISDA, Cutout, Mixup, and CutMix.

Furthermore, we explored the application of UMAC with multiple datasets (UMAC-MD), leveraging pre-set tasks and unsupervised pretraining to enhance model training. The UMAC-MD approach not only improved performance metrics but also demonstrated consistent results across different datasets. This highlights the practical applicability of the UMAC method, showing that its benefits extend beyond theoretical constructs to tangible improvements in real-world data scenarios.

Additionally, UMAC provides a structured approach to machine learning operations, offering a clear computation graph that details the flow of data and processing steps. This structured methodology enhances the understanding of how different components and algorithms interact within the model, ensuring a well-organized and consistent machine learning process.

Overall, this study illustrates the practical application of UMAC, confirming that its utility is not confined to theoretical exploration but can be effectively translated into improved performance in diverse and challenging datasets. The structured approach to machine learning operations provided by UMAC further contributes to its reliability and effectiveness, making it a valuable tool in medical image classification and other complex domains.

Author Contributions: Conceptualization, E.N. and Y.L.; methodology, E.N.; software, E.N.; validation, E.N. and Y.L.; formal analysis, E.N.; investigation, E.N.; resources, Y.L.; data curation, E.N.; writing—original draft preparation, E.N.; writing—review and editing, E.N. and Y.L.; visualization, E.N.; supervision, Y.L.; project administration, Y.L.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: E.N. and Y.L. declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACC	accuracy
AUC	area under the ROC curve
BSDA	Bayesian Random Semantic Data Augmentation
BYOL	Bootstrap Your Own Latent
CLSA	Contrastive Learning with Stronger Augmentations
CNN	Convolutional Neural Networks
EMA	Exponential Moving Average
EMA	Exponential Moving Average
ISDA	Implicit Semantic Data Augmentation
InfoNCE	Noise Contrastive Estimation
MLP	Multilayer Perceptron
MoCo	Momentum Contrast
MoCov2	Momentum Contrast Version 2
SOTA	State-of-the-Art
SimCLR	Simple Framework for Contrastive Learning
SimCLRv2	Simple Framework for Contrastive Learning Version 2
UMAC	Unified Model Agnostic Computation
ViT	Vision Transformer
XAI	Explainable Artificial Intelligence

References

1. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.; van Ginneken, B.; Sanchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
2. Shorten, C.; Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
3. Perez, L.; Wang, J. The effectiveness of data augmentation in image classification using deep learning. *arXiv* **2017**, arXiv:1712.04621.
4. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. AutoAugment: Learning Augmentation Strategies from Data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 113–123.
5. Ratner, A.; Bach, S.H.; Ehrenberg, H.; Fries, J.; Wu, S.; Ré, C. Snorkel: Rapid training data creation with weak supervision. In Proceedings of the VLDBEndowment, Munich, Germany, 28 August–1 September 2017; Volume 11, pp. 269–282.
6. Wang, S.; Jiang, L.; Shao, Z.; Sun, C.; Jia, J. Implicit semantic data augmentation for deep networks. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32, pp. 12632–12641.
7. Kim, J.; Park, J.; Shin, J.H.; Lee, J. PuzzleMix: Exploiting Saliency and Local Statistics for Optimal Mixup. In Proceedings of the International Conference on Machine Learning (ICML), Virtual Event, 13–18 July 2020.
8. Shahab Uddin, A.F.M.; Monira, S.; Monira, S.; Chung, T.C.; Bae, S.-H. SaliencyMix: A Saliency Guided Data Augmentation Strategy for Better Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 3–7 May 2021.
9. Qin, J.; Fang, J.; Zhang, Q.; Liu, W.; Wang, X.; Wang, X. ResizeMix: Mixing Data with Preserved Object Information and True Labels. *arXiv* **2021**, arXiv:2012.11101.
10. Chou, H.-P.; Chang, S.-C.; Pan, J.-Y.; Wei, W.; Juan, D.C. ReMix: Consistent and Adaptive Data Augmentation for Improved Generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Seattle, WA, USA, 14–19 June 2020.
11. Kim, J.H.; Choo, W.; Jeong, H.; Song, H.O. Co-Mixup: Saliency Guided Joint Mixup with Supermodular Diversity. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 3–7 May 2021.
12. Neghawi, E.; Liu, Y. Enhancing Self-Supervised Learning through Explainable Artificial Intelligence Mechanisms: A Computational Analysis. *Big Data Cogn. Comput.* **2024**, *8*, 58. [CrossRef]
13. Shen, L.; Margolies, L.R.; Rothstein, J.H.; Fluder, E.; McBride, R.; Sieh, W. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Sci. Rep.* **2017**, *7*, 244. [CrossRef]
14. McDermott, M.B.A.; Wang, S.; Marinsek, N.; Ranganath, R.; Foschini, L.; Ghassemi, M. Reproducibility in Machine Learning for Health. *Nat. Biomed. Eng.* **2021**, *5*, 1–2.
15. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
16. Oakden-Rayner, L. Exploring Large-scale Public Medical Image Datasets. *Acad. Radiol.* **2020**, *27*, 147–151. [CrossRef]
17. Kaushal, A.; Altman, R.B.; Langlotz, C.P. Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. *JAMA* **2020**, *324*, 936–938. [CrossRef]
18. Roberts, M.; Driggs, D.; Thorpe, M.; Gilbey, J.; Yeung, M.; Ursprung, S.; Aviles-Rivero, A.I.; Etmann, C.; McCague, C.; Beer, L.; et al. Common Pitfalls and Recommendations for Using Machine Learning to Detect and Prognosticate for COVID-19 Using Chest Radiographs and CT Scans. *Nat. Mach. Intell.* **2021**, *3*, 199–217. [CrossRef]
19. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. *Nature* **2017**, *542*, 115–118. [CrossRef] [PubMed]
20. Willemink, M.J.; Koszek, W.A.; Hardell, C.; Wu, J.; Fleischmann, D.; Harvey, H.; Folio, L.R.; Summers, R.M.; Rubin, D.L.; Lungren, M.P. Preparing Medical Imaging Data for Machine Learning. *Radiology* **2020**, *295*, 4–15. [CrossRef] [PubMed]
21. Cirillo, D.; Catuara-Solarz, S.; Morey, C.; Guney, E.; Subirats, L.; Mellino, S.; Gigante, A.; Valencia, A.; Rementeria, M.J.; Chadha, A.S.; et al. Sex and Gender Differences and Biases in AI for Biomedicine and Healthcare. *NPJ Digit. Med.* **2020**, *3*, 81. [CrossRef] [PubMed]
22. Vayena, E.; Blasimme, A.; Cohen, I.G. Machine Learning in Medicine: Addressing Ethical Challenges. *PLoS Med.* **2018**, *15*, e1002689. [CrossRef]
23. Raghupathi, W.; Raghupathi, V. Big Data Analytics in Healthcare: Promise and Potential. *Health Inf. Sci. Syst.* **2014**, *2*, 3. [CrossRef]
24. Johnson, A.E.; Pollard, T.J.; Shen, L.; Lehman, L.W.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; Mark, R.G. MIMIC-III, a Freely Accessible Critical Care Database. *Sci. Data* **2016**, *3*, 160035. [CrossRef]
25. Shurrab, S.; Duwairi, R. Self-Supervised Learning Methods and Applications in Medical Imaging Analysis: A Survey. *arXiv* **2021**, arXiv:2109.08685. [CrossRef]
26. Taleb, A.; Loetzsch, W.; Danz, N.; Severin, J.; Gaertner, T.; Bergner, B.; Lippert, C. 3D Self-Supervised Learning for Medical Imaging. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18157–18168.
27. Jamaludin, A.; Kadir, T.; Zisserman, A. Self-supervised learning for spinal MRIs. In Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Québec City, QC, Canada, 14 September 2017; pp. 294–302.

28. Li, H.; Fan, Y.H. Non-rigid Image Registration using Self-Supervised Fully Convolutional Networks without Training Data. In Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI), Washington, DC, USA, 4–7 April 2018; pp. 1075–1078.
29. Ribeiro, M.T.; Singh, S.; Guestrin, C. Why should I trust you? Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
30. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.
31. Droste, R.; Cai, Y.; Sharma, H.; Chatelain, P.; Drukker, L.; Papageorghiou, A.T.; Noble, J.A. Ultrasound Image Representation Learning by Modeling Sonographer Visual Attention. In *Lecture Notes in Computer Science*; Springer International Publishing: New York, NY, USA, 2019; pp. 592–604. [CrossRef]
32. Zhang, H.; Yang, J.; Gong, C.; Tao, D. Saliency-Guided Mixup. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 4007–4015. [CrossRef]
33. Wang, Y.; Huang, G.; Song, S.; Pan, X.; Xia, Y.; Wu, C. Regularizing Deep Networks with Semantic Data Augmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3733–3748 [CrossRef]
34. Zhu, Y.; Cai, X.; Wang, X.; Chen, X.; Yao, Y.; Fu, Z. BSDA: Bayesian Random Semantic Data Augmentation for Medical Image Classification. *arXiv* **2024**, arXiv:2403.06138.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
36. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500. [CrossRef]
37. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July; pp. 4700–4708. [CrossRef]
38. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
39. Yang, J.; Shi, R.; Wei, D.; Liu, Z.; Zhao, L.; Ke, B.; Pfister, H.; Ni, B. MedMNIST v2—A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Sci. Data* **2023**, *10*, 41. [CrossRef] [PubMed]
40. Gulrajani, I.; Lopez-Paz, D. In search of lost domain generalization. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 30 April 2020.
41. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
42. DeVries, T.; Taylor, G.W. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv* **2017**, arXiv:1708.04552.
43. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. *arXiv* **2017**, arXiv:1710.09412.
44. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Reoublic of Korea, 27 October–2 November 2019; pp. 6022–6031. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.