Enhancing Testicular Ultrasound Image Classification Through Synthetic Data and Pretraining Strategies

Nicola Morelli, Kevin Marchesini, Luca Lumetti, Daniele Santi, Costantino Grana, and Federico Bolelli ✉

University of Modena and Reggio Emilia, Italy
{*name.surname*}@unimore.it

Abstract. Testicular ultrasound imaging is vital for assessing male infertility, with testicular inhomogeneity serving as a key biomarker. However, subjective interpretation and the scarcity of publicly available datasets pose challenges to automated classification. In this study, we explore supervised and unsupervised pretraining strategies using a ResNet-based architecture, supplemented by diffusion-based generative models to synthesize realistic ultrasound images. Our results demonstrate that pretraining significantly enhances classification performance compared to training from scratch, and synthetic data can effectively substitute real images in the pretraining process, alleviating data-sharing constraints. These methods offer promising advancements toward robust, clinically valuable automated analysis of male infertility. The source code is publicly available at `https://github.com/AImageLab-zip/TesticulUS/`.

Keywords: Ultrasound · Medical Imaging · Synthetic · Diffusion Models

1 Introduction

Testicular UltraSound (TUS) imaging is a key non-invasive tool for evaluating male reproductive health by assessing tissue characteristics, such as parenchymal inhomogeneity, an emerging biomarker for male infertility [28]. However, subjective image interpretation and complex tissue patterns hinder reliable, standardized assessment, highlighting the need for automated classification tools.

Progress in this area is hampered by the lack of large, publicly available datasets. Ethical and privacy concerns limit data sharing, resulting in small, institution-specific datasets that constrain deep learning model development and hinder generalization. To address data scarcity, medical imaging research increasingly leverages pretraining [7] and synthetic data generation [25]. Supervised and self-supervised pretraining on large datasets enhances feature extraction and improves performance on smaller target datasets [18, 23]. Meanwhile, diffusion-based generative models [13] have emerged as powerful tools for synthesizing high-quality, realistic medical images, outperforming Generative Adversarial Networks (GANs) in stability and details modeling [24].

✉ Corresponding author: `federico.bolelli@unimore.it`

In this work, we evaluate supervised and self-supervised pretraining strategies for classifying testicular inhomogeneities using a ResNet-18 backbone. To address label noise common in ultrasound data, we propose a heuristic filtering method to improve training quality. Additionally, we explore diffusion-based synthetic data as a practical alternative to real images [4, 29], aiming to replicate the real data distribution and overcome data scarcity. Together, these strategies target improvements in both data quality and availability.

To summarize, our key contributions are: (*i*) a systematic evaluation of pretraining in testicular ultrasound analysis, (*ii*) a heuristic approach to reduce label noise, and (*iii*) the application of diffusion models for synthetic data generation in this sensitive, data-scarce domain.

2 Related Work

Deep Learning in UltraSound Image Analysis. Ultrasound (US) imaging is essential in medical diagnostics due to its safety, accessibility, and real-time capabilities, but interpretation remains challenging due to artifacts, noise, low contrast, and operator dependency [14]. Deep learning, particularly Convolutional Neural Networks (CNNs), has shown promise in automating analysis and extracting quantitative information [11, 17, 19, 20, 26].

However, US imaging presents unique challenges compared to modalities like MRI [21] or CT [3]. Limited annotated datasets, stemming from time-consuming, expert-dependent labeling, hinder model training, while heterogeneity and variability across devices and operators complicate generalization [30]. Privacy concerns further restrict data availability, impeding the development of robust models for applications like TUS analysis [27].

Generative Models for Synthetic Medical Image Generation. Generative models, particularly Generative Adversarial Networks (GANs) [10], have been used to alleviate data scarcity by augmenting datasets, performing cross-modality synthesis, and enabling anonymization [9, 15, 25], though they often suffer from training instability and limited diversity [2].

Recently, Denoising Diffusion Probabilistic Models (DDPMs) [13] emerged as a more effective alternative, achieving superior performance in generating realistic, diverse samples for MRI [24] and CT [22]. Conditional DDPMs further allow controlled generation based on clinical attributes or segmentation maps [29].

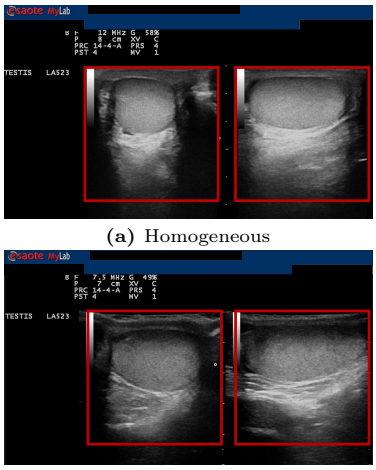
In this work, we address the underexplored domain of TUS classification by introducing the first benchmark targeting testicular pathology classification. To overcome data-sharing constraints, we demonstrate the effectiveness of DDPM-generated synthetic datasets when integrated into our pretraining pipeline.

3 Dataset Curation and Filtering

To the best of our knowledge, there is currently no publicly available dataset of testicular ultrasound images. Existing automatic approaches are primarily focused on testicular segmentation, and they typically rely on private datasets for training and evaluation [1]. For this reason, all the experiments presented

in this paper are conducted on an in-house dataset collected at the Antonio Nalin Center of the Baggiovara Hospital in Modena, Italy, using two different ultrasound acquisition systems: Esaote® MyLab25 Gold and Esaote® MyLab XPro80.[1]

The dataset includes image pairs, as illustrated in Fig. 1. Each pair contains static views of the same testicle, captured from transverse and sagittal planes. Pairs are cropped to remove metadata and isolate single views. Each view is treated separately, inheriting the original label.



(a) Homogeneous

(b) Inhomogeneous

Fig. 1: Example images provided by the clinical center. Red boxes indicate the regions selected during the cropping process.

Unlabeled Dataset (UD). While the primary focus of this work is on predicting the homogeneity versus inhomogeneity of testicular tissue, the dataset is enriched with thyroid ultrasound images, which are leveraged for pretraining purposes. Among the total 25 792 images, 1 664 correspond to testicular scans and 24 126 to thyroid scans, not necessarily from the same patients.

Labeled Dataset (LD). Additionally, for a subset of 880 testicular images, belonging to 220 patients, the inhomogeneity/homogeneity label is available, with a class distribution of approximately 20-80%. A significant challenge encountered during this project is the inherent noisiness in the pairing of images and labels. Ultrasound examinations are inherently dynamic, with clinicians relying on real-time video evaluation to assess anatomical properties. However, only static screenshots are saved during clinical practice. As a result, these images may not always accurately reflect the actual homogeneity characteristics of the tissue, introducing noise into the dataset.

Filtering Noisy Labels. To address this, we first developed an automatic filtering procedure applied to the 880 labeled images. This step aimed to identify and discard low-quality or misleading samples, thereby improving the reliability of the subsequent analyses and model training.

Leveraging a three-fold cross-validation schema, we train a simple ResNet-18 classifier for homogeneity classification based on the cross-entropy loss. Images from the same patient are always placed in the same fold to avoid data leakage. Results demonstrate poor classification performance and overfitting on training data. Particularly, it was clear that some of the examples were strongly perturbing the loss, indicating possible inconsistent labeling. A simple yet effective filtration schema has been adopted as follows. During model training, the per-sample loss values were recorded across epochs. Upon analyzing the loss trajectories, it was observed that most samples exhibited a near-monotonically

[1] Data will be anonymized and publicly released after receiving approval from the ethical committee. Download link: `https://ditto.ing.unimore.it/testiculus`.

decreasing loss trend. However, a subset of samples displayed highly irregular behavior, with spikes where the loss exceeded the value of 1. Empirically, a sample was flagged as "suspicious" if its training loss exceeded the threshold value of 1 on at least three[2] occasions during training. This evaluation process was repeated two times, leveraging ResNet-18 initialized with distinct pretrained weights, i.e., ImageNet and those provided by Chen *et al.* [7]. For each model, training was conducted using four random seeds and a three-fold cross-validation schema, resulting in a total of 24 runs. Due to the three-fold setup, each sample could be evaluated as "suspicious" between 0 and 16 times (i.e., appearing in multiple folds and seeds). Samples that were consistently flagged as "suspicious" in all 16 evaluations (72 images in total) were either discarded from the dataset or the corresponding label was flipped. Tab. 1 shows an improvement in model classification performance, confirming our hypothesis.

Table 1: ResNet-18 pretrained on ImageNet and fine-tuned on the *complete* LD dataset, *filtering* 72 "suspicious" images, or *flipping* their labels.

Dataset	Accuracy (↑)	F1-Score (↑)
Complete	81.51 ± 2.78	55.72 ± 4.37
Filtered	**88.15 ± 1.94**	68.59 ± 3.30
Flipped	86.78 ± 2.21	**73.17 ± 1.55**

Finally, a clinical evaluation was also performed. Clinicians were tasked to re-evaluate the labels of "suspicious" cases, this time using only the static views provided with the dataset since no video from the medical visit is available. Surprisingly, the evaluation was inconsistent with the original annotation, meaning that further study should rely not only on static images but on the entire video of the examination. All the experiments discussed in the rest of the paper were performed with the dataset resulting from such a polishing operation.

4 Methods

This section describes the strategies we propose to pretrain the classification model in a semi-supervised fashion and the process we leverage for generating and filtering synthetic ultrasound images. We also detail the neural network architectures evaluated, the fine-tuning procedure on the target classification task, and the evaluation protocol adopted to assess the synthetic data generation.

4.1 Pretraining Strategies

For the classification task, we selected a ResNet-18 architecture [11] as the backbone model.[3] It is widely recognized that, in such low-data regimes, models can benefit from pretraining strategies that enable better feature extraction [7]. Therefore, we explored effective approaches for pretraining the network to enhance its performance on the classification task (Fig. 2).

In contrast with classical ImageNet-based pretraining or other existing approaches [7], we investigated two different sources of data for pretraining:

– Real ultrasound images of the thyroid and testicular areas, using our UD dataset described in Sec. 3;

[2] Thresholds of four and five exceedances were also tested, but found to be less effective.

[3] Preliminary experiments showed that more complex architectures, such as ResNet-50 and Vision Transformers, tended to overfit, given the limited size of our LD dataset.

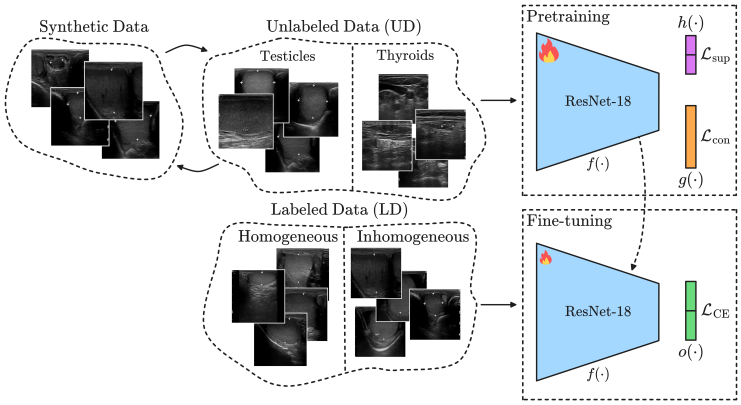


Fig. 2: Proposed pretraining leveraging synthetic or UD, and fine-tuning on the LD.

– Synthetic ultrasound images of testicles, generated using a diffusion model with a procedure to filter out-of-distribution samples, detailed in Sec. 4.2.

For the pretraining task, we employed a semi-supervised approach combining contrastive learning with supervised classification of the type of organ targeted in the ultrasound image (thyroid or testicle in our case). Specifically, we used the SimCLR framework [6] for the unsupervised contrastive component and a cross-entropy loss for the classification of the organ.

Contrastive Pretraining. SimCLR is a contrastive learning framework aimed at training an image encoder $f(\cdot)$ to produce representations that are invariant to image augmentations. This is achieved by maximizing the agreement between differently augmented views of the same image (positive pairs), while minimizing the agreement between views of different images (negative pairs).

Given a batch of N images $\{x_i\}_{i=1}^{N}$, we apply a stochastic augmentation pipeline twice to each image, resulting in two correlated views $(\tilde{x}_{2i-1}, \tilde{x}_{2i})$ per image. This effectively yields a batch of $2N$ augmented examples. Each augmented view $\tilde{x}_k$ is passed through a shared encoder $f(\cdot)$ followed by a projection head $g(\cdot)$, resulting in projected representations $z_k = g(f(\tilde{x}_k))$.

The similarity $\mathrm{s}(z_j, z_k)$ between any pair of representations is measured using cosine similarity, and the contrastive loss $\ell(j,k)$ for a positive pair is defined via a normalized temperature-scaled cross-entropy, as follows:

$$\mathrm{s}(z_j, z_k) = \frac{z_j^\top z_k}{\|z_j\| \, \|z_k\|}, \qquad \ell(j,k) = -\log \frac{\exp\left(\mathrm{s}(z_j, z_k)/\tau\right)}{\sum_{m=1}^{2N} \mathbb{1}_{[m \neq j]} \exp\left(\mathrm{s}(z_j, z_m)/\tau\right)}, \tag{1}$$

where τ is a temperature hyperparameter, and $\mathbb{1}_{[m \neq j]}$ is an indicator function equal to 1 when $m \neq j$, and 0 otherwise. Then the contrastive loss is computed by averaging over all positive pairs in the batch:

$$\mathcal{L}_{\text{con}} = \frac{1}{2N} \sum_{i=1}^{N} \left[\ell(2i-1, 2i) + \ell(2i, 2i-1)\right]. \tag{2}$$

Supervised Classification. To further enhance the learned representations, we incorporate a supervised classification objective during pretraining. For this

purpose, each image is annotated with a label corresponding to its anatomical region, i.e., *thyroid* or *testicle*, and a classification head $h(\cdot)$ is attached to the encoder $f(\cdot)$ to predict these labels. For each augmented view, we predicted the logits $c_k = h(f(\tilde{x}_k))$, and the supervised loss is computed using the cross-entropy across all pairs of augmented views:

$$\mathcal{L}_{\text{sup}} = \frac{1}{2N} \sum_{i=1}^{N} \big(\text{CE}(c_{2i-1}, y_i) + \text{CE}(c_{2i}, y_i) \big). \tag{3}$$

The final pretraining objective is a weighted combination of the contrastive and supervised losses:

$$\mathcal{L} = \mathcal{L}_{\text{con}} + \lambda \, \mathcal{L}_{\text{sup}}, \tag{4}$$

where λ is set to 0.2 to balance the contribution of the supervised loss.

4.2 Synthetic Data Generation and Filtering

To address data scarcity in ultrasound imaging and overcome privacy-related data sharing constraints, we explored synthetic image generation for pretraining. A Denoising Diffusion Probabilistic Model (DDPM) was used to produce high-quality synthetic images as a substitute for real data. Specifically, we used the framework introduced by [8], which has demonstrated superior performance over GANs in image synthesis tasks.

The diffusion model operates through a two-phase process: a forward diffusion phase and a reverse denoising phase. In the forward phase, Gaussian noise is incrementally added to an image over multiple time steps, transforming a clean image into pure noise. This process is defined by a Markov chain, where each step adds a small amount of noise, controlled by a predefined variance schedule. In the reverse phase, a U-Net architecture is trained to reconstruct the original image by progressively removing the added noise. The model learns to predict the noise component and the diagonal covariance matrix of the noise distribution at each time step, allowing it to denoise the image iteratively. During inference, starting from a Gaussian noise sample, the model iteratively refines this noise through a series of denoising steps. At each step t, the model estimates the noise component $\epsilon_\theta(x_t, t)$ present in the current noisy image and computes a less noisy image x_{t-1}. This iterative process continues until the final step $t = 0$, resulting in a synthetic image x_0 that resembles the distribution of real ultrasound images. For our application, we trained the diffusion model on the LD dataset of real testicular ultrasound images.

Evaluation Metrics for Synthetic Images. We employed three established metrics to evaluate the quality of the generated images: improved *precision* and *recall*, both introduced by Kynkäänniemi *et al.* [16], and the *Fréchet Inception Distance* (FID) [12]. The precision assesses the fidelity of the generated images, quantifying the distributional similarity between real and generated data, while the recall measures the diversity of the synthetic data, indicating how much of the real data distribution is covered by the synthetic samples. FID, by comparing

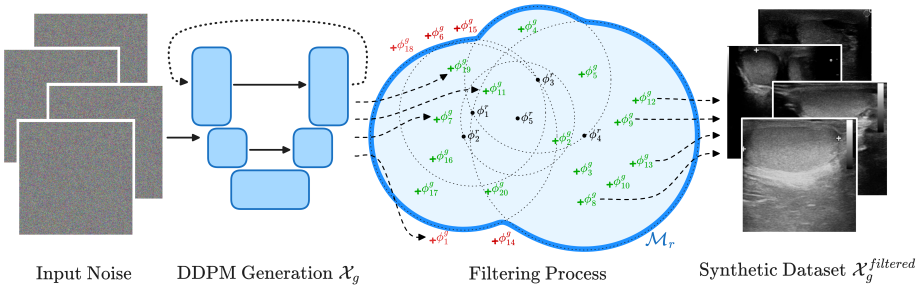


Fig. 3: Overview of the pipeline used for synthetic data generation and filtering.

both the mean and the covariance of the real and generated feature distributions, captures both two aspects.

The process to compute these metrics involves embedding both real and generated images into a high-dimensional feature space using a pretrained network (i.e., Inception). Let $\Phi_r = \{\phi_1^r, \phi_2^r, \ldots, \phi_N^r\}$ denotes the set of feature vectors for real images, and $\Phi_g = \{\phi_1^g, \phi_2^g, \ldots, \phi_M^g\}$ for generated images. For each real image feature vector ϕ_i^r we define an hypersphere $B(\phi_i^r, r_i)$ centered at ϕ_i^r, where the radius is the distance to its k-th nearest neighbor in Φ_r (symmetrically hyperspheres $B(\phi_j^g, r_j)$ are constructed around each generated sample ϕ_j^g using its k-th nearest neighbor in Φ_g).

Defining the real data manifold $\mathcal{M}_r$ (respectively, the generated data manifold $\mathcal{M}_g$) as the union of all the real data (generated data) hyperspheres:

$$\mathcal{M}_r = \bigcup_{i=1}^{N} B(\phi_i^r, r_i), \qquad \left(\mathcal{M}_g = \bigcup_{j=1}^{M} B(\phi_j^g, r_j)\right), \tag{5}$$

the precision P is computed as the fraction of generated samples whose embeddings fall inside the real data manifold $\mathcal{M}_r$, and the recall R is the fraction of real samples falling inside the generated data manifold $\mathcal{M}_g$:

$$P = \frac{1}{M}\sum_{j=1}^{M} \mathbb{1}_{\mathcal{M}_r}(\phi_j^g), \qquad R = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\mathcal{M}_g}(\phi_i^r), \tag{6}$$

where $\mathbb{1}_{\mathcal{M}_r}(\phi_j^g)$ (respectively $\mathbb{1}_{\mathcal{M}_g}(\phi_i^r)$) is an indicator function which is 1 if $\phi_j^g \in \mathcal{M}_r$ ($\phi_i^r \in \mathcal{M}_g$), 0 otherwise.

The FID assumes that the feature vectors of real and generated data, extracted from the pretrained Inception network, follow multivariate Gaussian distributions. Let μ_r and Σ_r be the mean and covariance of the real image features Φ_r, and μ_g and Σ_g those of the generated image features Φ_g. The FID is defined as the Fréchet distance between these two distributions:

$$\mathrm{FID} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{1/2}\right). \tag{7}$$

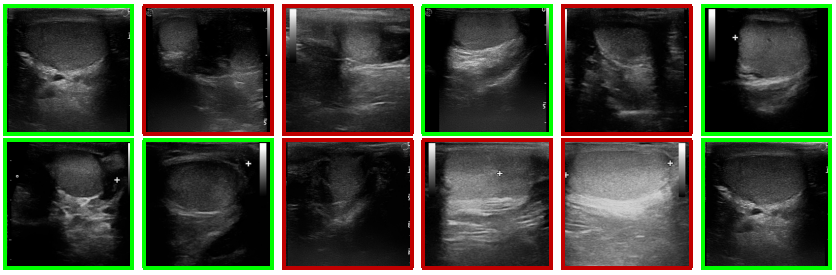


Fig. 4: Sample images from our generated dataset. Green represents high-quality samples, while red identifies those discarded by our filtering method.

A lower FID value indicates that the generated images are statistically more similar to the real ones, reflecting both high fidelity and appropriate variability in the synthetic data distribution.

Algorithm 1 Filtering Algorithm.

Input: $\mathcal{X}_g$, Φ_r, Φ_g, # of neighbors k.
Initialize $\mathcal{X}_g^{\text{filtered}} \leftarrow \emptyset$
for all ϕ_i^r in Φ_r **do**
  Compute $r_i \leftarrow$ Euclidean distance to the k-th nearest neighbor of ϕ_i^r in Φ_r
  $B(\phi_i^r, r_i) \leftarrow$ hypersphere centered at ϕ_i^r with radius r_i
end for
$\mathcal{M}_r \leftarrow \bigcup_{i=1}^{N} B(\phi_i^r, r_i)$
for all ϕ_j^g in Φ_g **do**
  if $\phi_j^g \in \mathcal{M}_r$ **then**
    Add x_j^g to $\mathcal{X}_g^{\text{filtered}}$
  end if
end for
return $\mathcal{X}_g^{\text{filtered}}$

Filtering Method. To ensure the quality of the generated synthetic ultrasound images, we developed a filtering method based on the abovementioned precision metric. Specifically, we computed the real data manifold, $\mathcal{M}_r$, from the feature embeddings of the LD dataset used to train the diffusion model, as specified in Eq. (5). Given the set of generated images $\mathcal{X}_g = \{x_1^g, x_2^g, \ldots, x_M^g\}$, we compute their corresponding feature representations, Φ_g, and filter them by selecting only those whose embeddings lie inside $\mathcal{M}_r$, obtaining the final filtered synthetic dataset $\mathcal{X}_g^{\text{filtered}}$ (Algorithm 1, Fig. 3).

Generation Results. Following [8], to compute precision and recall metrics on generated data, we set the number of neighbors $k = 3$ and leverage three non-overlapping reference batches of 64 real images sampled from the LD dataset. Instead, since the filtering process is based on the entire LD dataset, it leverages a $k = 50$. We found that k should scale approximately linearly with the number of real data employed in the computation of the manifold to maintain hyperspheres of comparable size across different settings.

Applying our filtering, we increased the precision of the synthetic dataset from 79.68 ± 4.81 to 90.1 ± 4.70. As a natural consequence, some generated samples were removed, leading to a reduction of the recall from 25.0 ± 1.56 to 11.9 ± 3.25. However, the variations in recall remained limited, and the FID decreased from 119.84 ± 2.39 to 116.84 ± 4.12, confirming that the filtering strategy achieved a good compromise between maintaining similarity to the real data distribution and preserving adequate coverage of the feature space. After filtering, the original $\sim 20K$ synthetic samples were reduced to $\sim 9K$.[4] Samples of generated images are available in Fig. 4.

[4] Filtered synthetic data are publicly released at `https://ditto.ing.unimore.it/testiculus`.

Table 2: Three-fold cross-validation results for ResNet-18 on the homogeneous and inhomogeneous downstream task, starting from different pretraining strategies.

Pretraining	# Samples	Accuracy (↑)	F1-Score (↑)	Precision (↑)	Recall (↑)
∼	∼	73.93 ± 6.80	56.05 ± 7.11	45.48 ± 9.91	75.12 ± 3.66
ImageNet	1.28M	83.89 ± 2.15	67.89 ± 1.85	61.72 ± 3.08	76.08 ± 4.11
USCL [7]	23.00K	75.46 ± 2.64	57.43 ± 1.63	47.37 ± 2.16	73.86 ± 4.76
UD (only testicles)	1.66K	81.09 ± 2.95	65.44 ± 2.32	55.86 ± 2.74	79.55 ± 2.79
UD (only thyroids)	24.13K	80.13 ± 2.83	63.92 ± 2.46	54.30 ± 3.87	78.88 ± 2.06
UD	25.79K	**86.78 ± 2.21**	**73.17 ± 1.55**	**67.84 ± 1.67**	**80.27 ± 2.13**

4.3 Fine-tuning

Fine-tuning was conducted on the LD dataset using a standard transfer learning setup. The pretrained ResNet-18 backbone was used as a feature extractor, and a lightweight classification head $o(\cdot)$ (Fig. 2), consisting of a single linear layer, was appended on top to perform binary classification. To prevent the network from focusing on superficial visual cues, such as spurious patterns or annotations, we added random synthetic markers to each image (Fig. 5) during training. This forced the network to learn more robust and generalizable features. In addition to marker insertion, we applied a series of spatial data augmentations, including random rotations, horizontal flipping, and small random shifts, to further improve generalization and mitigate overfitting. Finally, in order to mitigate class imbalance, we used a weighted sampler.

5 Experiments and Results

Implementation Details. The ResNet-18 backbone was pretrained on synthetic ultrasound data using supervised and unsupervised objectives (batch size 1024) on two NVIDIA L40S GPUs (48 GB), with the LARS optimizer and a polynomial learning rate schedule (initial value 10^{-3}). Inputs were normalized (mean 0.5, std. dev. 0.25). Fine-tuning for binary classification employed three-fold cross-validation over four random seeds, using a batch size of 64 on a single RTX 2080 Ti GPU. The backbone and classification head were fine-tuned with learning rates of 10^{-5} and 10^{-4}, respectively. Synthetic ultrasound images were generated using a diffusion model retrained from scratch on domain-specific data (batch size 16, 256×256 input resolution) with a single L40S GPU, following [8]. Both mean and variance were learned, and sampling was unguided.

On the Role of Pretraining. To validate the effectiveness of the proposed pretraining strategy, we applied it to ResNet using our UD dataset. For comparative analysis, we also considered ResNet pretrained on ImageNet and the ultrasound-specific pretrained weights provided by Chen *et al.* [7], devised for ultrasound data, although focused on lung and liver. In each of the three pretraining scenarios, the model was subsequently fine-tuned on our LD dataset following a three-fold cross-validation schema. All data from the same patient were strictly confined to the same fold to effectively prevent data leakage. To ensure a fair comparison and reliable convergence, the number of training steps

Table 3: Three-fold results on the downstream task when pretraining ResNet-18 with different combinations of real and synthetic data with ($\mathcal{X}_g^f$) or without ($\mathcal{X}_g$) applying the proposed filtering procedure.

Real Testicle	Real Thyroids	Synthetic Testicle	Accuracy (↑)	F1-Score (↑)	Precision (↑)	Recall (↑)
✗	✗	$\mathcal{X}_g$	80.58 ± 3.58	64.24 ± 2.31	55.46 ± 3.08	77.13 ± 1.70
✗	✗	$\mathcal{X}_g^f$	80.42 ± 2.71	64.77 ± 1.62	54.69 ± 3.24	80.12 ± 2.02
✓	✗	∼	81.09 ± 2.95	65.44 ± 2.32	55.86 ± 2.74	79.55 ± 2.79
✗	✓	$\mathcal{X}_g$	82.88 ± 2.11	67.61 ± 1.56	58.87 ± 1.71	79.59 ± 3.63
✗	✓	$\mathcal{X}_g^f$	84.60 ± 2.08	70.63 ± 1.20	62.19 ± 1.76	**82.17 ± 0.97**
✓	✓	∼	**86.78 ± 2.21**	**73.17 ± 1.55**	**67.84 ± 1.67**	80.27 ± 2.13

was held constant across all experiments, regardless of variations in the size of the pretraining dataset. This ensures that each model undergoes the same total number of forward and backward passes. The results are summarized in Tab. 2.

Our initial observation indicates that leveraging ultrasound-specific pretrained weights from Chen *et al.* (third row of Tab. 2) does not necessarily yield optimal performance, particularly when the pretraining data originates from different ultrasound acquisition systems, as in the case of the USCL dataset [7]. Conversely, when images are sourced from the same acquisition system, variations in the anatomical regions, namely testicles and thyroid, did not significantly impact performance. Specifically, downstream performance on testicular imaging was nearly identical regardless of whether testicular or thyroid images were employed during pretraining (fourth and fifth rows of Tab. 2). It is important to highlight that in these instances, pretraining leveraged exclusively the contrastive component of the loss in Eq. (4). Interestingly, the ImageNet-pretrained model achieved the best results, highlighting the superior generalization of features learned through supervised training on diverse natural images, even across domains as different as natural and ultrasound images.

Finally, combining ultrasound data from different anatomical structures, in this case testicular and thyroid images (last row of Tab. 2), enabled the integration of supervised and unsupervised losses, delivering the best overall performance. This strategy resulted in a noticeable improvement, increasing accuracy by approximately 3 points and the F1-score by about 6 points compared to the ImageNet pretrained baseline. An ablation study to compare the impact of loss weightings is also conducted in Tab. 4. Results show both losses affect performance, with $\lambda = 0.2$ yielding the best outcome.

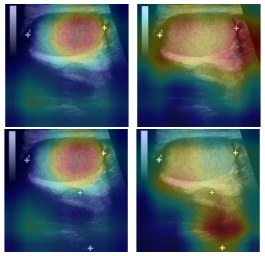
The Impact of Synthetic Data. Tab. 3 demonstrates that using filtered synthetic data ($\mathcal{X}_g^{\text{filtered}}$) consistently yields better performance compared to unfiltered synthetic data ($\mathcal{X}_g$). Although the highest accuracy is achieved when both real datasets are included, the performance of models trained solely on synthetic data remains competitive, exhibiting only a modest degradation. These findings underscore the practical utility of synthetic data,

Table 4: Ablation study on using different loss components and varying λ, Eq. (4).

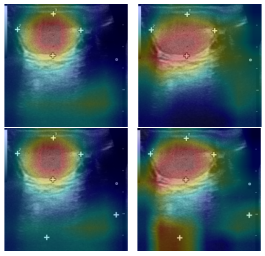
$\mathcal{L}_{\text{con}}$	$\mathcal{L}_{\text{sup}}$	λ	Accuracy (↑)	F1-Score (↑)
✓	✗	0.0	84.77 ± 2.56	69.41 ± 0.51
✓	✓	0.1	85.38 ± 2.71	71.40 ± 0.93
✓	✓	0.2	**86.78 ± 2.21**	**73.17 ± 1.55**
✓	✓	0.3	85.29 ± 2.68	71.36 ± 0.70
✓	✓	0.4	85.03 ± 2.43	71.22 ± 1.26
✗	✓	1.0	83.88 ± 2.56	70.11 ± 1.56

particularly in scenarios where access to real data is restricted due to privacy concerns or availability limitations.

Qualitative Evaluation. A heatmap visualization generated using Grad-CAM++ [5] on two representative samples from the test set of our LD dataset is reported in Fig. 5. For each sample, four images are shown: outputs of the same model trained either **with** (left column) or **without** (right column) applying the augmentation strategies reported in Sec. 4.3. The second row for each sample differs from the first by the introduction of artificial markers ✣. As can be observed, the use of augmentation strategies helped focus the model's attention more precisely on the inner part of the testicle, the region most closely associated with the inhomogeneity property relevant to our downstream task. Furthermore, the proposed augmentation approach, which introduces synthetic random artifacts during training, proved effective in mitigating the influence of such artifacts on the predictions. This effect is particularly evident when comparing the bottom-left and bottom-right images of each sample.



(a) Sample 1.



(b) Sample 2.

Fig. 5: Grad-CAM++.

6 Conclusion and Future Research Directions

In this study, we addressed key challenges in the automated classification of testicular ultrasound inhomogeneity, a promising biomarker for male infertility. By combining supervised and unsupervised pretraining with diffusion-based synthetic augmentation, we achieved significant improvements over models trained from scratch or using pretraining strategies tailored for ultrasound imaging.

Future work will focus on incorporating dynamic ultrasound videos, which may offer richer contextual information for classification. We also aim to develop label-conditioned synthetic image generation to produce datasets suitable for both pretraining and fine-tuning. Advancing these directions will be essential for the next generation of automated, clinically deployable tools.

Acknowledgements. This project is funded by the University of Modena and Reggio Emilia and Fondazione di Modena through FAR-2024 (E93C24002080007) and FARD-2024, and by the Italian Ministry of Research's NRRP complementary actions "Fit4MedRob – Fit for Medical Robotics" (PNC0000007).

References

1. Abdalla, A.M., et al.: Automatic Segmentation and Detection System for Varicocele Using Ultrasound Images. CMC **72**(1) (2022)
2. Arora, S., et al.: Generalization and Equilibrium in Generative Adversarial Nets (GANs). In: ICML (2017)

3. Bolelli, F., et al.: Segmenting Maxillofacial Structures in CBCT Volumes. In: CVPR (2025)
4. Cartella, G., et al.: Unveiling the Truth: Exploring Human Gaze Patterns in Fake Images. IEEE SPL (2024)
5. Chattopadhay, A., et al.: Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In: WACV (2018)
6. Chen, T., et al.: A Simple Framework for Contrastive Learning of Visual Representations. In: ICML (2020)
7. Chen, Y., et al.: USCL: Pretraining Deep Ultrasound Image Diagnosis Model Through Video Contrastive Representation Learning. In: MICCAI (2021)
8. Dhariwal, P., et al.: Diffusion Models Beat Gans on Image Synthesis. In: NeurIPS (2021)
9. Frid-Adar, M., et al.: GAN-based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion Classification. Neurocomputing (2018)
10. Goodfellow, I.J., et al.: Generative Adversarial Nets. In: NeurIPS (2014)
11. He, K., et al.: Deep Residual Learning for Image Recognition. In: CVPR (2016)
12. Heusel, M., et al.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: NeurIPS (2017)
13. Ho, J., et al.: Denoising Diffusion Probabilistic Models. In: NeurIPS (2020)
14. Huang, Q., et al.: Machine Learning in Ultrasound Computer-Aided Diagnostic Systems: A Survey. BioMed Research International **2018**(1) (2018)
15. Kazeminia, S., et al.: GANs for Medical Image Analysis. AIME **109** (2020)
16. Kynkäänniemi, et al.: Improved Precision and Recall Metric for Assessing Generative Models. In: NeurIPS (2019)
17. Litjens, G., et al.: A survey on deep learning in medical image analysis. MedIA **42** (2017)
18. Lumetti, L., et al.: U-Net Transplant: The Role of Pre-training for Model Merging in 3D Medical Segmentation. In: MICCAI
19. Lumetti, L., et al.: Enhancing Patch-Based Learning for the Segmentation of the Mandibular Canal. IEEE Access (2024)
20. Lumetti, L., et al.: Taming Mambas for 3D Medical Image Segmentation. IEEE Access (2025)
21. Menze, B.H., et al.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE TMI **34**(10) (2014)
22. Pan, S., et al.: Synthetic CT Generation from MRI Using 3D Transformer-Based Denoising Diffusion Model. Medical Physics **51**(4) (2024)
23. Panariello, A., et al.: Consistency-Based Self-supervised Learning for Temporal Anomaly Localization. In: ECCV (2022)
24. Pinaya, W.H., et al.: Brain Imaging Generation with Latent Diffusion Models. In: MICCAI Workshop (2022)
25. Pollastri, F., et al.: Augmenting Data with GANs to Segment Melanoma Skin Lesions. MTAP **79**(21-22) (2019)
26. Porrello, A., et al.: Spotting Insects from Satellites: Modeling the Presence of Culicoides Imicola Through Deep CNNs. In: SITIS (2019)
27. Price, W.N., et al.: Privacy in the age of medical big data. Nature Medicine **25**(1) (2019)
28. Spaggiari, G., et al.: Testicular ultrasound inhomogeneity is an informative parameter for fertility evaluation. AJA **22**(3) (2020)
29. Wang, W., et al.: Semantic image Synthesis Via Diffusion Models. ArXiv Preprint arXiv:2207.00050 (2022)
30. Yap, M.H., et al.: Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks. IEEE J-BHI **22**(4) (2017)