Mammo-LeJEPA: A Data-Efficient Self-Supervised Framework for Breast Density Estimation

Ricardo Montoya-del-Angel, Enric Sena-Alvarez, Eloy Garcia, Hadeel Awwad, Joan Marti, and Robert Marti

Computer Vision and Robotics Institute, University of Girona, Girona, Spain
ricardo.montoya@udg.edu

Abstract. Automated breast density estimation from mammography is crucial for breast cancer screening, but developing robust deep learning baselines is often hindered by limited medical data volume and representation collapse in self-supervised regimes. In this paper, we introduce Mammo-LeJEPA, a data-efficient self-supervised framework that explicitly models the structural symmetry of screening mammography. Rather than relying on large-scale natural image pretraining or complex engineering heuristics to prevent collapse, our method exploits native multi-view constraints by mapping paired cranio-caudal (CC) and mediolateral oblique (MLO) projections through a shared Vision Transformer encoder. We restrict the latent-space geometry using a non-parametric Sliced Isotropic Gaussian Regularization loss to extract invariant clinical morphology while preventing representation collapse. Mammo-LeJEPA is trained entirely from scratch on the public VinDr-Mammo dataset. In downstream evaluations using frozen linear probes, our compact embeddings achieve a binary density estimation AUROC of 0.9754, outperforming a much larger ViT-Large JEPA alternative, and maintain a robust One-vs-Rest AUROC of 0.9410 on the challenging 4-class BI-RADS task without downstream fine-tuning. The code is available at https://github.com/IMPACT-project-UdG/Mammo_LeJEPA.

Keywords: LeJEPA · mammography · breast density · self-supervised learning

1 Introduction

The paradigm of deep learning is rapidly shifting toward foundation models, particularly world models designed to learn by understanding the underlying physics and structure of their environment rather than relying on massive, data-hungry reconstruction tasks. In this context, self-supervised learning (SSL) architectures can be broadly divided into generative/reconstruction models and joint-embedding predictive architectures (JEPA) [1]. While generative approaches, such as latent diffusion [7], successfully project inputs into a latent space, they optimize for pixel-level reconstruction through a compressed representation. Consequently, they retain sensitivity to high-frequency pixel noise and low-level

artifacts. In contrast, JEPA bypasses low-level noise entirely by operating strictly within the embedding space. By breaking an input into distinct views and predicting the embedding of one view from another, JEPA captures high-level semantics, encouraging the latent space to map global clinical morphology and general structural biomarkers.

In breast imaging, the clinical target is not faithful reconstruction but the extraction of stable patient-specific tissue patterns. Breast density estimation is a global assessment of the relative amount of fibroglandular and fatty tissue across the breast, routinely derived from screening mammography. The paired CC and MLO views are complementary observations of the same anatomy, so a model that agrees across both views is encouraged to suppress projection artifacts and retain the information that matters for breast density estimation.

The semantic robustness of JEPA has prompted its recent application to medical imaging, where capturing structural biomarkers without background noise is critical. In breast imaging, AEGIS [9] and BAC-JEPA [10] show the promise of joint-embedding architectures, but they still rely on hand-crafted heuristics to avoid representation collapse.

LeJEPA [2] addresses this issue with a simple inductive bias: constraining the latent representation space into an isotropic Gaussian distribution. This prevents collapse while also reducing the amount of data required for training, making in-domain pretraining more practical for medical cohorts.

We propose a novel deep learning framework based on the LeJEPA architecture, tailored for breast density estimation. Rather than relying solely on artificial augmentations, our model exploits the native CC/MLO pairing available in standard mammography to build a symmetric view-to-view embedding objective. We validate the data efficiency of our framework by training directly on the limited data provided by the VinDr dataset [4].

The key contributions of our work are threefold. First, we introduce Mammo-LeJEPA, a novel self-supervised learning framework for mammography analysis based on the LeJEPA architecture that effectively prevents representation collapse without relying on complex engineering heuristics. Second, we propose a symmetric view-to-view embedding architecture that leverages corresponding CC and MLO mammogram views to capture comprehensive high-level structural biomarkers. Finally, we show that in-domain pretraining from scratch on a moderately sized dataset can yield competitive representations, providing a lightweight framework for localized medical imaging cohorts.

2 Methodology

2.1 Framework Overview: Symmetric View-to-View Representation Learning

The fundamental premise of a Joint-Embedding Predictive Architecture (JEPA) involves mapping a context input and a target input into a shared latent space via independent encoders, followed by a prediction network that infers the target

embedding from the context embedding [1]. In the proposed framework, we adapt this paradigm for breast density estimation by leveraging the inherently multi-view nature of mammography screening. We utilize the paired Cranio-Caudal (CC) and Medio-Lateral Oblique (MLO) projections of the same breast.

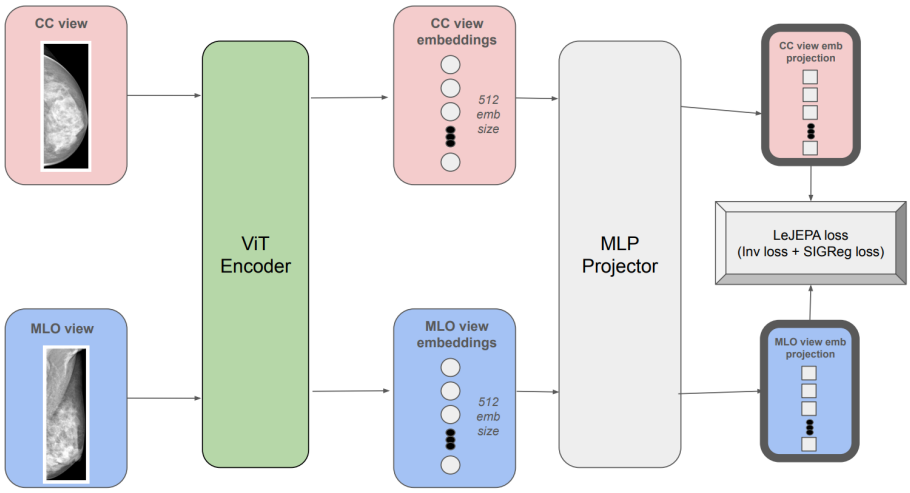

Fig. 1. LeJEPA mammography pipeline overview. The paired CC and MLO views are processed through a shared ViT encoder. Then, the embeddings are projected into a projected space where the invariance loss and SIGReg regularization are applied.

Let v_1 and v_2 represent these anatomically symmetric views, CC and MLO in our case. Because there is no inherent information asymmetry between these paired views (i.e., they represent symmetric global views rather than masked versus unmasked regions), an additional predictor network is redundant [2]. Thanks to LeJEPA's SIGReg loss, we can remove both the predictor and teacher-student architecture without suffering from representation collapse. The independent encoders thus merge into a single, shared siamese ViT encoder, as shown in Figure 1. Consequently, the explicit prediction mapping collapses into an identity function, and the predictability criterion is accomplished by directly comparing the embeddings of the views with a metric. The objective thus shifts to ensuring that the embedding representation remains structurally invariant across the different mammographic projections. In the case of LeJEPA, this is achieved by minimizing the Mean Squared Error (MSE) between the projected embeddings of the two views, while simultaneously enforcing a Gaussian inductive bias on the latent space to prevent collapse.

2.2 Network Architecture

The proposed LeJEPA architecture consists of a shared backbone encoder, a non-linear projector, and an online diagnostic probe.

– **Encoder (Backbone):** We utilize a Vision Transformer (ViT) as the core backbone, parameterized by θ. Specifically, a ViT-Small architecture with a patch size of 16 is adapted for a targeted mammography resolution of 512×512. Given the two input views v_1 and v_2, the shared encoder extracts high-dimensional representations $h_1 = f_\theta(v_1)$ and $h_2 = f_\theta(v_2)$. The ViT architecture effectively captures the global spatial dependencies of the breast morphology.

– **Projector Network:** To prevent the backbone from collapsing into simple trivial invariances and to map the features into the space where the constraints will be applied, the representations are passed through a projector network g_ϕ. The projector acts as an information filter, implemented as a Multi-Layer Perceptron (MLP) mapping the 512-dimensional backbone features through 2048-dimensional hidden layers to the final projection dimension, regularized via 1D Batch Normalization. The projected embeddings are defined as $z_1 = g_\phi(h_1)$ and $z_2 = g_\phi(h_2)$.

– **Online Diagnostic Probe:** To monitor the representation quality of the downstream task during pretraining without affecting the JEPA backpropagation, we introduce an independent, parallel probe network. This online diagnostic tool takes the detached representations h_1 and h_2 (acting as frozen features) and monitors performance strictly on the binary breast density estimation task (low vs. high density). Configured structurally as a sequential layer normalization followed by a linear projection (`nn.Sequential(nn.LayerNorm(512), nn.Linear(512, 2))`), it dynamically tracks the binary classification metrics.

2.3 Objective Function: Invariance and SIGReg

The total objective of our LeJEPA framework aims to enforce embedding invariance across views while strictly regulating the latent space geometry to prevent representation collapse. This dynamically balances the two competing forces of self-supervised learning: pulling representations of the same image together while pushing representations of different images apart to fill the latent space uniformly.

Prediction (Invariance) Loss Due to the identity function of the prediction network in our symmetric view case, the predictive loss naturally simplifies to an invariance loss. We minimize the Mean Squared Error (MSE) between the projected embeddings of the CC and MLO views, forcing the network to discard projection-specific noise and retain global anatomical features:

$$\mathcal{L}_{inv} = \frac{1}{B} \sum_{i=1}^{B} \|z_1^{(i)} - z_2^{(i)}\|_2^2, \tag{1}$$

where B is the batch size. In practice, this acts as an elegant method for comparing multiple views by penalizing the squared deviation of the projections from their mean.

Sliced Isotropic Gaussian Regularization (SIGReg) To avoid representation collapse, LeJEPA applies a Gaussian inductive bias to the embedding space. However, due to the explosion of computation requirements needed for calculating full covariance matrices in high dimensions, we utilize Sliced Isotropic Gaussian Regularization (SIGReg). SIGReg is a mathematical trick that projects the high-dimensional embedding space onto random 1-dimensional vectors and matches these projections to a target 1D standard normal distribution $\mathcal{N}(0, 1)$ using statistical distance metrics.

Let $w \in \mathbb{S}^{d-1}$ be a random projection direction sampled uniformly from the unit sphere so that the 1D projection of our embeddings is $z^w = z \cdot w$. We minimize the Epps-Pulley test statistic T for normality of z^w against $\mathcal{N}(0, 1)$:

$$\mathcal{L}_{\text{SIGReg}}(z) = \mathbb{E}_{w \sim \mathbb{S}^{d-1}} \left[T(z \cdot w) \right]. \tag{2}$$

This allows us to efficiently compute the loss in a projected 1D Gaussian-constrained space while globally regulating the latent geometry.

The total pretraining loss $\mathcal{L}_{\text{LeJEPA}}$ is defined as a weighted trade-off between the invariance loss and the SIGReg loss applied to the views, governed by a hyperparameter λ:

$$\mathcal{L}_{\text{LeJEPA}} = (1 - \lambda)\mathcal{L}_{inv} + \lambda \left(\mathcal{L}_{\text{SIGReg}}(z_1) + \mathcal{L}_{\text{SIGReg}}(z_2) \right). \tag{3}$$

The online diagnostic probe is evaluated using a standard supervised cross-entropy loss between the predicted outputs and the ground-truth labels. Because the encoder features are detached before feeding into the probe, the self-supervised and supervised losses remain independent.

3 Experiments

Dataset and Preprocessing Our experiments were conducted using the public VinDr-Mammo dataset [4]. The preprocessing pipeline was designed to standardize the heterogeneous mammogram acquisitions. Initially, pixel inversion was applied to MONOCHROME1 images to align their representation with MONOCHROME2 scans. To isolate the breast tissue, a region of interest cropping was performed by setting the global background to zero. This process utilized Otsu's thresholding and connected components analysis to clean the masks, particularly in cases affected by dark noise. Subsequently, robust min-max scaling was executed using percentiles calculated exclusively over the breast tissue pixels. In instances with significant dark noise, the lower percentile boundary was increased to prevent contrast degradation. Finally, the intensities were normalized to the [0,1] range and saved as 16-bit floating-point tensors to reduce the memory footprint while preserving radiometric fidelity. The paired Cranio-Caudal (CC) and Medio-Lateral

Oblique (MLO) projections for each breast were structured as symmetric inputs for the architecture. To prevent data leakage across views and studies, we used the patient-level data split provided by the original dataset authors, which splits the data into 80% for training and 20% for testing.

Implementation Details The feature extractor relies on a ViT-Small architecture with a patch size of 16, accepting images at a 512×512 resolution. The backbone was trained from scratch without prior initialization and incorporates a drop path rate of 0.1. The projector acts as an information filter, implemented as a Multi-Layer Perceptron (MLP) mapping the 512-dimensional backbone features through two 2048-dimensional hidden layers before the final projection layer, regularized via 1D Batch Normalization. The pretraining backbone (ViT-Small) contains 21.7M parameters, while the projector network adds 5.51M parameters, yielding a total pretraining size of 27.2M parameters.

The full training process was separated into 2 stages: the self-supervised pretraining and the downstream evaluation. The first stage, the self-supervised pretraining, was conducted solely using the VinDr-Mammo training split. Due to the training paradigm of self-supervised learning, pretraining hyperparameters, including the loss trade-off parameter $\lambda = 0.02$, initial learning rate 5×10^{-4}, and the 600-epoch training schedule, were fixed *a priori* based on standard self-supervised learning protocols [2] without hyperparameter search or exposure to downstream task labels. This fixed-budget strategy guarantees complete isolation of the self-supervised pretraining phase from downstream diagnostic evaluations. To generate symmetric views during training, data augmentation pipelines were applied, consisting of random horizontal flipping, random affine transformations (rotations up to 8 degrees and 5% translations), randomized center jitter square cropping scaled between 0.7 and 1.0, and a random intensity gain control multiplier ranging from 0.9 to 1.1. Deterministic center cropping without jitter or scaling was utilized during the evaluation phase.

An online diagnostic probe was used only for real-time tracking of the downstream task performance during pretraining, not for model selection or hyperparameter tuning. It was structured strictly for the binary density classification task (low density [BI-RADS A–B] vs. high density [BI-RADS C–D]), defined as `nn.Sequential(nn.LayerNorm(512), nn.Linear(512, 2))`.

For the second stage, a comprehensive downstream evaluation was performed using the frozen backbone features, evaluating on both the binary configuration and a full 4-class layout (`nn.Linear(512, 4)`). The frozen features of the training set were used to train a supervised offline linear probe for both the binary and 4-class breast density downstream task. For similar reasons as the pretraining stage, no hyperparameter search was conducted for the downstream evaluation and a fixed number of 600 epochs, batch size of 256 and a learning rate of 3×10^{-3} were used for all downstream linear probes.

To evaluate the influence of the latent bottleneck space in the downstream task performance, an ablation study was conducted, where the projection dimension was set to 16, 32, 64, and 128. We trained on an NVIDIA A40 GPU, using 41

GB of VRAM with a batch size of 112 for the LeJEPA model, and 314 MB for the downstream linear probes. The training time for the LeJEPA model over 600 epochs was approximately 24 hours, and for the downstream linear probes approximately 15 minutes.

Evaluation Metrics To assess model performance and effectively address potential class distribution imbalances across the dataset, we utilized four distinct metrics for downstream breast density classification evaluation: Accuracy, Balanced Accuracy, Macro F_1-score, and Area Under the Receiver Operating Characteristic Curve (AUROC).

3.1 Results

Pretraining Convergence Analysis In the context of the proposed self-supervised architecture, the pretraining loss trajectories primarily serve as empirical validation of optimization stability and the prevention of representation collapse. Our analysis assessed the training dynamics across the four different projection bottleneck configurations (i.e., 16, 32, 64, and 128-dimensional). Observation of the anticollapse regularization indicated that the magnitude of the SIGReg penalty scaled proportionally with the dimensionality of the bottleneck. This behavior is an expected mathematical artifact rather than a reflection of underlying representation quality. Because the SIGReg objective calculates statistical deviations across the axes of the projection head, higher-dimensional vectors inherently possess a larger capacity for dimensional variance, which naturally translates into higher raw regularization magnitudes.

On the other hand, the trajectory of the invariance loss remained consistent across all experimental configurations. The convergence curves for the 16, 32, 64, and 128-dimensional bottlenecks followed nearly identical optimization paths. This result demonstrates that the degree of compression applied at the final projection layer does not impede the capacity of the Vision Transformer to geometrically align the symmetric CC and MLO views. The shared backbone successfully learns to extract the core structural invariants of the breast morphology entirely independent of the tightness of the projection bottleneck.

Regarding the online probe, the classification loss exhibited a stable convergence pattern across all bottleneck dimensions. For all four experiments, the AUROC curve converged to values around 0.97, indicating that the learned representations are highly informative for the downstream breast density classification task. The stability of the probe's performance across varying bottleneck sizes further reinforces the robustness of the LeJEPA framework in capturing clinically relevant features without succumbing to representation collapse.

Breast Density Classification Performance The downstream linear probing performance across differing tasks and projection dimensions is summarized in Table 1. The framework exhibits strong semantic compression: expanding the projection dimension from $proj\_dim = 16$ to 128 yields only marginal

performance shifts across all metrics. In the binary task, the lower-dimensional bottleneck achieves an virtually identical AUROC of 0.9754 compared to 0.9751 at dimension 128. This indicates that the combination of our symmetric view-invariance constraint and the non-parametric anticollapse regularizer (SIGReg) effectively packs the essential structural features of breast tissue distribution into highly compact representations without spreading signal across redundant latent axes. When compared against standard binary baselines that use the same split structures [8], our lightweight backbone remains competitive. Crucially, the AEGIS framework [9] requires a much larger ViT-Large backbone to achieve a binary classification AUROC of 0.9545 on the VinDr-Mammo cohort. In contrast, our framework achieves 0.9754 using a smaller ViT-Small encoder, supporting the value of the view-to-view symmetric inductive bias.

Table 1. Downstream linear probing performance on VinDr-Mammo embeddings at 600 epochs across differing tasks and projection dimensions.

Configuration	Accuracy	Balanced Acc.	Macro F_1	AUROC
Binary Task				
LeJEPA ($proj\_dim = 16$)	0.9025	0.9236	0.8020	0.9754
LeJEPA ($proj\_dim = 128$)	0.9055	0.9275	0.8069	0.9751
4-Class Task				
LeJEPA ($proj\_dim = 16$)	0.7535	0.6718	0.5600	0.9403
LeJEPA ($proj\_dim = 128$)	0.7540	0.6864	0.5604	0.9410

In the multi-class ordinal setting, the model maintains a high threshold-independent discriminative capacity with a One-vs-Rest (OvR) AUROC of 0.9410 ($proj\_dim = 128$), though threshold-dependent metrics naturally adjust (Accuracy = 0.7540, Balanced Acc. = 0.6864, Macro F_1 = 0.5604). This drop stems from the continuous biological spectrum of breast tissue making adjacent boundaries highly ambiguous, alongside the severe minority class scarcity of Categories A and D within the cohort. Contextualizing these multiclass results within the current SOTA landscape highlights LeJEPA's distinct advantages in reproducibility and simplicity. While multi-view architectures like [5] struggle with generalization, failing to present direct VinDr test benchmarks and dropping to a Macro F_1 of 0.3544 on external verification sets, our frozen features retain structural stability. Meanwhile, the Attention-Guided Erasing (AGE) method [6] reaches a slightly higher Macro F_1 score of 0.5910 ($\pm$0.017) on the 4-class task. However, direct comparison highlights the efficiency and simplicity of our architecture. While AGE relies on out-of-domain initialization via an ImageNet-pretrained DeiT backbone [6], followed by an involved multi-stage downstream protocol, including empirical attention-head selection heuristics, dynamic binary mask generation, and full end-to-end backbone fine-tuning, our framework operates strictly within the localized data boundaries. Our linear probe achieves a Macro F_1 score of 0.5604 using purely frozen, unaugmented

embeddings. This removes downstream image-space modifications, prevents backpropagation through the heavy transformer backbone during downstream training, and reduces operational complexity and inference overhead. Lastly, while the evidential framework [3] reports an exceptional Macro AUROC of 0.9892 ($\pm$0.007), their model requires heavy pretraining across four independent mammography datasets. LeJEPA approaches this performance envelope while training strictly from scratch within the localized boundaries of the VinDr cohort. Taken together, these comparisons demonstrate that view-to-view symmetric LeJEPA yields a clinically representative, lightweight, and effective embedding space optimal for downstream diagnostic translation.

4 Conclusion

Mammo-LeJEPA demonstrates that a symmetric view-to-view joint-embedding predictive architecture can extract highly representative, robust mammographic features without relying on large-scale supervised initialization or cumbersome anti-collapse heuristics. By leveraging the native geometric pairing of screening examinations, the model achieves strong diagnostic performance on both binary and 4-class breast density tasks without requiring external pretraining or large-scale supervised initialization. These findings demonstrate that explicitly modeling multi-view clinical constraints offers a streamlined, lightweight, and accurate paradigm for future medical foundation models. While our model demonstrates strong representation capability, several design choices and limitations warrant explicit acknowledgement. First, our framework utilizes native CC and MLO views as an intuitive anatomical prior without an internal control ablation against single-view synthetic augmentations. Additionally, controlled data-volume scaling, multi-center external validation, and direct head-to-head benchmarking against established self-supervised methods under identical split conditions were beyond the scope of this study. Finally, while the Epps-Pulley test statistic within SIGReg effectively prevented representation collapse, evaluating alternative statistical test statistics for latent space regularization remains an open area for future investigation.

Acknowledgments. This research was possible thanks to funding from the IMPACT project (PID2024-157201OB-C21) from the Ministerio de Ciencia, Innovación y Universidades of Spain.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15619–15629 (2023)

2. Balestriero, R., LeCun, Y.: Lejepa: Provable and scalable self-supervised learning without the heuristics. arXiv preprint arXiv:2511.08544 (2025)
3. Gudhe, R., Mazen, S., Sund, R., Kosma, V.M., Behravan, H., Mannermaa, A.: An evidential deep learning framework for assessment of mammograms. Research Square (11 2023). https://doi.org/10.21203/rs.3.rs-2821378/v1
4. Nguyen, H.T., Nguyen, H.Q., Pham, H.H., Lam, K., Le, L.T., Dao, M., Vu, V.: Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. Scientific Data **10**(1), 277 (2023)
5. Nguyen, H.T., Tran, S.B., Nguyen, D.B., Pham, H.H., Nguyen, H.Q.: A novel multi-view deep learning approach for bi-rads and density assessment of mammograms. In: 2022 44th Annual international conference of the IEEE engineering in medicine & biology society (EMBC). pp. 2144–2148. IEEE (2022)
6. Panambur, A.B., Yu, H., Bhat, S., Madhu, P., Bayer, S., Maier, A.: Attention-guided erasing: Novel augmentation method for enhancing downstream breast density classification. In: BVM Workshop. pp. 13–18. Springer (2024)
7. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
8. Sharifian, P., Hong, X., Karimian, A., Amini, M., Arabi, H.: Attention-enhanced deep learning ensemble for breast density classification in mammography. arXiv preprint arXiv:2507.06410 (2025)
9. Waggener, S.C., Navuluru, S.K., Tamil, L.: Aegis: A multi-task joint-embedding predictive architecture for mammography. arXiv preprint arXiv:2607.00277 (2026)
10. Waggener, S.C., Tamil, L.: Bac-jepa: Label-efficient breast arterial calcification segmentation via synthetic mammography-guided supervision. arXiv preprint arXiv:2606.22089 (2026)