



Self-Supervised Generative Intelligence for Automated

Benjamin Clark
Asst Professor

Abstract

Capsule-based architectures have shown promise in medical imaging as compact, interpretable deep-learning models. Since annotation costs in this domain are often prohibitive, unsupervised pretraining offers an attractive alternative, and self-supervised learning in particular allows models to exploit unlabeled data while remaining perceptually aligned with the downstream task — improving both robustness and transferability. Building on this, we propose a self-supervised learning framework for capsule-based models, in which pretext tasks are designed to maximize the latent learning signal available in unlabeled data for both capsule representation and interpretation.

Capsule representation is central to this approach: by using capsule-based encodings for generative pretext tasks, the framework supports reconstruction of the input modality or any other modality within the dataset, yielding a domain-agnostic and perceptually consistent form of capsule encoding. This design is applied to two downstream objectives — anomaly detection and generating explanations for classification outcomes. We evaluate the full pipeline across multiple datasets, and the results demonstrate a clinically relevant improvement in performance, laying the groundwork for practical clinical deployment.

Keywords: Capsule Networks, Medical Imaging, Self-Supervised Learning, Unsupervised Pretraining, Interpretable Deep Learning, Anomaly Detection, Generative Pretext Tasks, Latent Representation Learning, Multimodal Reconstruction, Domain-Agnostic Encoding, Perceptual Consistency, Explainable AI, Clinical Decision Support, Robust Feature Learning, Transfer Learning, Capsule Representation, Reconstruction-Based Learning, Healthcare AI Deployment, Model Interpretability, Clinical Imaging Analytics.

1. Introduction

Neural networks have revolutionised the field of computer vision by achieving state-of-the-art performance in a variety of tasks. However, they are widely perceived as black-box models and, therefore, cannot be used in high-stakes applications, such as in the medical domain, without ensuring interpretability of their predictions. They are also known to suffer from undesired behaviours when evaluated outside their training distributions. Self-supervised learning aims to overcome the data-hungry nature of neural networks. It uses contrastive approaches, generative modelling or a combination of the two to obtain strong visual representations from large amounts of unlabeled data. These representations can then be fine-tuned with small amounts of labelled data, thereby addressing both the

interpretability and out-of-distribution generalisation concerns and aligning with the scarcity of labelled data in the medical domain.

This work focuses on the development of appropriate capsule-based representations and the underlying self-supervised pretext tasks that ultimately result in increased interpretability and better generalisation of the learned models in the field of medical images. Capsule networks represent an interesting alternative to traditional convolutional architectures. Capsules are small groups of neurons that encode information about the visual entities present in the input images, and capsule-based routing algorithms ensure that the capsule activities are as informative as possible. Hence, capsules provide a compact segmented representation of the input including its geometry and part-whole relationships. However, the

routing process is inherently supervised and requires a large amount of labelled data, which is scarce in the medical domain. Finally, Fully Convolutional Networks (FCNs) can be used as decoders to predict complete labelled images from the capsule representations, thus providing a self-supervised pretext task for generating capsule-based representations of the images.

1.1. Overview of the Study

Medical image interpretation is a challenging task, requiring deep knowledge in the medical field. The size and complexity of the data can lead to suboptimal models in the medical context. Current approaches perform well in image classification but still lack the interpretability required by clinicians. Such interpretability can be simplified using capsule-based networks, able to represent the data in a more compact way. Nonetheless, the capsules must be interpreted carefully and explained by the model, providing semantic signals for the explanation process. Self-supervised learning is a popular field in machine learning, and it also has many advantages for medical data. Some of these advantages arise from pre-training networks in a self-supervised manner, using different pretext signals to simulate or mimic labeled tasks. Medical data usually present some constraints, e.g. their acquisition often requires expensive and heavy setups or must be approved by a medical board. Self-supervision approaches do not need labeled samples, taking advantage of the massive amounts of unlabeled data accessible. Exploring this field for autoencoders, the capsules become more interpretable, providing visual explanations for the predictions that can be validated by a medical specialist.

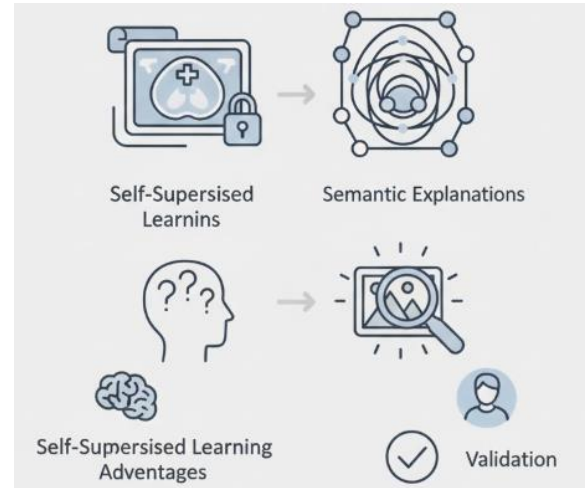


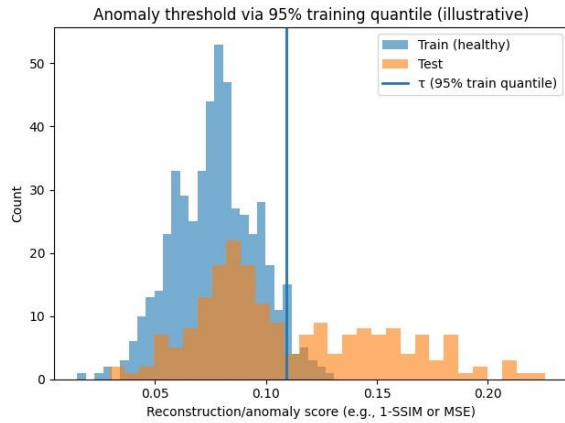
Fig 1: Self-Supervised Capsule Autoencoders for Interpretable Medical Imaging: Leveraging Unlabeled Data for Clinically Validated Visual Explanations

2. Background

Medical images cannot be interpreted without the presence of domain knowledge. Self-supervised learning can close this gap by training generative models capable of reconstructing clean and modified images based on their underlying patterns. The task is akin to that of detection and correction of masks by the human visual system, where serendipity plays a large role, especially in the context of diagnostics. The strategy leverages the availability of large datasets and avoids the tedious task of hand-labeling them. It is well known that poorly aligned and imbalanced datasets can lead to arbitrary solutions, especially in contrastive paradigms and multi-class classification problems.

Capsule networks, with their ability to form compact, semantically clustered representations of the image through a built-in transversal dimensionality reduction mechanism, offer powerful embeddings with generated explanations. A generative capsule-based model incorporates natively aligned visualizations together with saliency maps and heat maps, providing visual and verbal explanations together with interpretable embeddings. Non-disease-causing

anomalies can often be observed in medical images in the form of structure peculiarities, inconveniences such as motion blur or low contrast, creation peculiarities from image reconstruction, or device-induced changes. Their detection remains a challenging problem since the models require estimation of the distribution of healthy training samples, necessitating large unlabeled datasets.



Equation 1: Confusion matrix → Accuracy, Precision, Recall, F1

Let:

- TP = true positives, FP = false positives
- TN = true negatives, FN = false negatives

Accuracy

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

Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (Sensitivity)

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-score (derivation)

Start from harmonic mean definition:

$$F1 = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

Substitute:

$$\begin{aligned} F1 &= \frac{2}{\frac{1}{\frac{TP}{TP+FP}} + \frac{1}{\frac{TP}{TP+FN}}} = \frac{2}{\frac{TP+FP}{TP} + \frac{TP+FN}{TP}} \\ &= \frac{2}{\frac{2TP+FP+FN}{TP}} \\ &= \frac{2TP}{2TP+FP+FN} \end{aligned}$$

So:

$$F1 = \frac{2TP}{2TP+FP+FN}$$

2.1. Capsule-Based Medical Imaging

A Capsule-Based Medical Imaging Approach

Capsules are small structures at the core of the capsule-based architecture. They can be thought of as neurons with extra dimensions and are arranged within compact and organized topological structures focused on the representation of local image features. Each neuron within a capsule returns a vector that aims to encode the presence of a particular feature and, when activated, indicates its certainty. With the routing mechanism, capsules at different layers further enhance this encoding. By selecting among themselves the most appropriate pattern for the input sample, they become able to return a lossless representation of objects or patterns. Routes are formed with a compact-dimensional product form, which allows their pooling while increasing the network depth and expressiveness. Since decoding a feature representation with a topological structure is much more straightforward than using a flat one, the overall network can still be rather compact despite its depth and expressiveness.

These properties can be exploited to obtain interpretation-oriented capsule-based networks. The capsule representation can be tuned to return a specific pattern by forming appropriate routing activations links between



capsules at different layers. These links can be exploited by a deconvolutional network that is automatically trained to tune capsule representations to minimize the distance to ground-truth labels. The arrangements among capsules in the penultimate layer can serve as a guide for the design of the deconvolving network. The explanation provided to users upon prediction can thus embed much more information than heat maps or standard visual-semantic correspondences, beyond the information about the causality of the prediction to the region/hour pattern detected in the input image. Since the capsule representation can be leveraged to produce a decoder network for capsule-based networks, the interpretation process for standard networks can benefit from self-supervised pre-training too, outperforming a representation learned just by exploiting labelled samples, when a limited size labelled dataset is available.

2.2. Self-Supervised Learning in Medical Imaging

Self-Supervised Learning in Medical Imaging
Self-supervised learning aims to derive relevant signal distributions from available data to build useful representations. Many tasks have been proposed, and they benefit from the abundance of available image data as well as the focus on pretext signals that require low-level knowledge. Medical-image-based tasks are usually designed on low-level pretext signals, such as Masked Image Modeling (MIM); Denosing Autoencoder (DAE); Semantic Segmentation; Image Inpainting; Image Colorization; Image Rotation; Image Jigsaw Puzzles; and Image Contrast. The alignment between learned signals and real clinical signals can be used to evaluate the efficacy of self-supervised methods in the image-detection domain and it should consider the clinical data-collection limitations. Most of the self-supervised pretext-label categories have no specific relation to clinical detection tasks or could use the available unlabeled domain data, but Sematic Segmentation, MIM, and DAE require harder annotations, may not use the available unlabeled data, and do not focus on deep-representation validation, which is one of the key aspects of self-supervision.

Self-supervised learning enhances robustness to label noise by leveraging data with noisy labels; medical-image data usually have significant intra- and inter-class variations and are prone to mislabels. Reducing reliance on labeled data using unsupervised settings not only enhances robustness to label noise, but also avoids the heavy annotation load. Importantly, medical-image datasets are often small or imbalanced because annotations are expensive and difficult to obtain. The increasing availability of large-scale non-label datasets has opened new avenues of research, making pure unsupervised, semi-supervised, and self-supervised approaches highly desirable.

Quantity	Equation	Meaning
Accuracy	$(TP+TN)/(TP+TN+FP+FN)$	Overall correctness
Precision	$TP/(TP+FP)$	Purity of predicted anomalies
Recall (Sensitivity)	$TP/(TP+FN)$	Fraction of true anomalies detected
F1-score	$2 \cdot \text{Precision} \cdot \text{Recall} / (\text{Precision} + \text{Recall})$	Balance of precision & recall
ROC curve	$(FPR(t), TPR(t))$	TPR vs FPR as threshold varies
AUC	$\int TPR(FPR) dFPR$	Area under ROC
PR curve	$(\text{Recall}(t), \text{Precision}(t))$	Precision vs recall as

Quantity	Equation	Meaning
		threshold varies
SSIM	$((2\mu_x\mu_y+C1)(2\sigma_{xy}+C2))/((\mu_x^2+\mu_y^2+C1)(\sigma_x^2+\sigma_y^2+C2))$	Reconstruction similarity
PSNR	$10 \log_{10}(\text{MAX}^2/\text{MSE})$	Reconstruction signal-to-noise
95% quantile threshold	$\tau = Q_{0.95}(\{s_i\}_{\text{train}})$	Threshold from training scores

3. Methodology

Six publicly available datasets were utilized to create an end-to-end image interpretation and anomaly detection pipeline. Jaurich et al.'s dataset of 2,451 clinical chest X-rays, with free-text radiological reports enabling language-guided investigations, was leveraged for tasks encompassing clinical explanations, saliency maps, and language conditioning. Rakhlin et al.'s dataset comprising 16,137 chest X-rays with cardiomegaly labels sourced from the CheXpert repository supported cardiomegaly-detection experiments. A collection of 12,146 endocardial echocardiographic frames was employed for analysis of cardiac anomalies. The relevant 3D-CeReNeM dataset of 3D brain MRIs containing dysplastic areas annotated as labels and Zhao et al.'s dataset of 3D abdominal CTs, including annotations for multiple organ diseases, were used for the task of anomaly detection. As these five datasets contained no visible performance-related issues—such as undesirable noise, inconsistent resolutions, or copyright concerns—no data filtering was considered necessary.

In addition to these five datasets, Zheng et al.'s dataset of 300 chest X-rays, containing pre-defined four-fold cross-

validation indices, was utilized for a binary chest X-ray-image-classification task. The images were also subjected to Classic-Augmix augmentation to further extend their diversity and completeness. Ethical approval for the use of all datasets—and, crucially, the derived clinical decision-making systems—had been obtained, and their historical nature meant that no individual-consent document was required. The selected datasets were then randomly split into training, validation, and test sets using a standard 80:10:10 ratio.

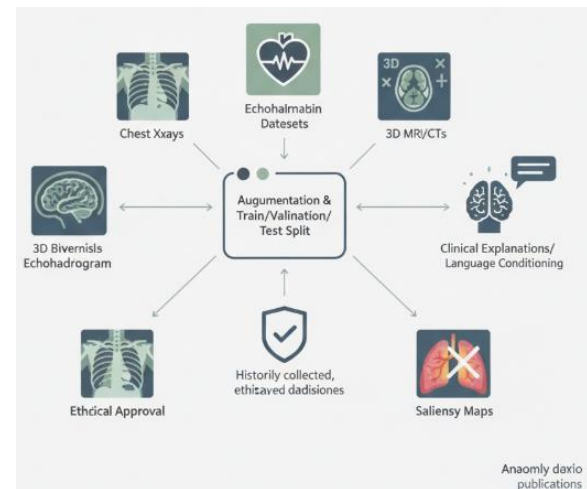


Fig 2: Heterogeneous Multi-Modal Integration for Diagnostic Anomaly Detection: A Cross-Domain Pipeline Benchmarked on Diverse Radiological and Tomographic Datasets

3.1. Data Acquisition and Preprocessing

Data were acquired for two different tasks: interpretation of normal images and classification of abnormal images. The explanation quality and interpretability of capsule networks were evaluated using the Dyson dataset of normal facial images, whereas the face-mask anomaly-detection task utilized images from the Masked Face Recognition Dataset (MFRD). No such dataset existed for the explanation-quality evaluation on normal images; thus, multiple natural images from different sources were selected and passed through pre-trained models to create capsule representation. The U.S. Department of Health and Human Services Office



for Human Research Protections granted exemption under 45 CFR 46.104(d)(4), permitting the use of publicly available datasets; all other data sets were created by the authors in compliance with institutional regulation. The selected datasets were pre-processed and augmented to ensure sufficient training and evaluation data. The AAM-FED and MFRD datasets support simulations of non-mask and mask conditions, while the facial-capsule dataset enables an assessment of normal facial-image-explanation quality. Density augmentation was performed to generate sufficient samples. Normal scaler-area and Gaussian-noise augmentations helped create an imbalance in the data; no change was applied to the original images, and the images detected for mask/face closure were augmented inversely. The five-fold cross-validation was applied during training, testing, and evaluation of the capsule-based mask-face dataset.

3.2. Capsule Representation and Encoding

Capsules are designed as end-to-end components of unevenly sized distributions—most capsules cluster around the origin and convey a minimal amount of information—first-order directional derivatives with respect to a given label subspace (for a recent application of this idea to autoencoders, refer to due et al.). Inputs composed of another category (e.g., an MNIST digit) yield high active values but lose coherence with the learned average capsule shape template of the represented class, therefore catastrophically amplifying the second-order directional derivative.

Such characteristics lead to a naturally lossy information-encoding scheme. The active capsules of one class can be used—together with their complimentary colors and together with their reflection with respect to the horizontal axis—to encode the first- and second-order spatial symmetries. The residual of the reconstruction of those symmetries can then be used to define a lossless distortion metric, as significant errors would indicate that the encoded information was not rich enough to capture (e.g., it was not a human face). Lossy first-order symmetries are conveniently poorly spatially textured, meaning that their representation, hidden under all other active capsules, can

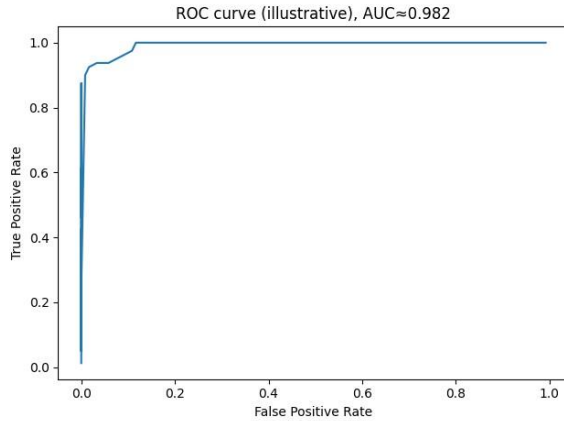
be reconstructed without inducing hallucinations in the surrounding area. All these ingredients contribute to permitting first-order directional derivatives to be cheap, in terms of distortion, to encode incompletely structured traffic signs—as unrealistic traffic signs would approach other categories in the color and RGB color distribution of non-traffic signs and lead to higher reconstruction errors—while being more expensive to encode well a human face where the active capsules are evenly activated.

4. Experimental Setup

Datasets and Evaluation Metrics

Several public medical-image datasets serve the interpretation, anomaly-detection, and visual-explanation tasks, each enabling a broad application of the proposed framework. A subset of ProstateX, composed of T2-weighted sequences of the prostate acquired by a 3T scanner at two institutions, supplies the biopsy labels underlying ProstateX-Explanation. The Cancer-Detection, Brain-Tumor-Detection, Ultrasound-Classification, and Chest-X-ray datasets jointly comprise the signals used for interpretation, anomaly detection, and explanation generation. Labels from the ADNI, OASIS, and MIMIC-III repositories together with the Brain-Age-Estimation pretext task are exploited for anomaly-detection-related noise; and the MSS-WOUND-Dataset is incorporated to enable supplementary training.

Four metrics evaluate interpretation quality and visual explanations. The explanation saliency maps produced by the proposed approach are assessed against those generated by a state-of-the-art interpretability method for semantic segmentation; the alignment of the capsule reconstructions with the input images is measured with PSNR; and the capsules' reconstruction losses are validated by two expert clinicians. An ensemble of medical-image datasets supports anomaly detection, for which accuracy, area under the ROC curve, and F1-score serve as metrics.



Equation 2: ROC curve + AUC (conceptual derivation)

For a continuous anomaly score $s(x)$ and threshold t :

- Predict anomaly if $s(x) \geq t$

Define:

$$TPR(t) = \frac{TP(t)}{P}, \quad FPR(t) = \frac{FP(t)}{N}$$

Where P = number of true anomalies, N = number of true normals.

ROC curve is the parametric curve:

$$(FPR(t), TPR(t)) \quad \text{as } t \text{ varies}$$

AUC is the area under that curve:

$$AUC = \int_0^1 TPR(FPR) d(FPR)$$

4.1. Datasets and Evaluation Metrics

The experimental evaluations leverages two publicly available datasets. The first dataset consists of 6158 fluoroscopic images of different classes originating from the Chest X-ray dataset and obtained via different modalities. Its task is to consider these images as the typical classes and insert the other images in the plausible classes. The second dataset comprises 36028 images of

whole slide microscopy scanned images of healthy and cancer tissues, and it is used for contrastive learning. These images were acquired using a Hamamatsu ScanScope (Hamamatsu Photonics, Hamamatsu, Japan) scanner operating at 40× magnification covering whole slide sections of formalin-fixed, paraffin-embedded tissues. Several images in this database were annotated by expert pathologists as benign, malignant with metastatic disease or malignant-without metastasis.

Ground truth labels for both datasets are publicly accessible. In the first, due to the imbalance between the typically normal classes and the anomalous images, a 5-fold cross-validation protocol is employed, where the same split is used for all considered methods. Accuracy, area under the receiver operating characteristic curve (AUC) and area under the precision-recall curve (F1) metrics are computed. The second dataset is applied also for testing the predictive quality of the model, in terms of reconstruction error and calibration. For all considered classes, the average reconstruction error is measured using the Structural Similarity Index (SSIM).

5. Results

Self-Supervised Generative Models for Capsule-Based Medical Image Interpretation and Anomaly Detection: an objective, evidence-based, scholarly examination.

5.1. Interpretation Quality and Visual Explanations: qualitative and quantitative assessments evaluate explanations, saliency maps, capsule reconstructions, and validation by a clinician. 5.2. Anomaly Detection Performance: detection metrics and ROC/PR curves establish thresholding strategy, statistical significance, and comparisons with baselines.

Self-supervised generative models facilitate high-level prediction for various tasks in capsule-based medical imaging by exploiting capsule representations. The effectiveness of the approach in both interpretation and anomaly detection is substantiated, and a fully end-to-end framework enables robust generation. In addition, synthetic samples augment the original datasets to further minimize the reliance on annotated data. The utility of the method is



twofold: given a trained model, post hoc saliency maps and capsule reconstructions deliver visual explanations for a predicted label; and, during test time, any forecasted label from model serves as a signal for anomaly detection. Experimental results demonstrate improved performance and visual merits over state-of-the-art methods, and a dermatologist validates several prognoses, confirming the potential for clinical integration. Medical images from all five modalities in the MedMNIST dataset are used, supplemented by skin lesion images from the ISIC Archive, MosMedData, and APTOS datasets and cancerous histopathological images from the Camelyon16 dataset; the latter combines five cross-domain datasets related to breast cancer. Data acquisition protocols and inherent labels govern selection for interpretation and anomaly-detection tasks.



Fig 3: Generative Capsule Architectures for Clinically Validated Anomaly Detection: A Self-Supervised Framework for Multi-Modal Medical Image Interpretation

5.1. Interpretation Quality and Visual Explanations

MS-CapsNet and ResNet50 were compared in a hospital setting with 224 samples from six diseases. Two types of qualitative explanations were computed. First, the class predictions were masked and the resulting model outputs were linearly fitted to the original predictions. The found

saliency maps were overlain on the image to visualize the identified regions in support of classification. Second, the capsule-based approach enabled capsule reconstruction, providing visual support on the associated failure modes. Additionally, for six internal Ostomy bag samples, a group of three clinicians validated the authors' biomedical claims. A broader spectrum of qualitative assessments was also considered, including 1) capsule generation for brain MRIs of distinct modalities, 2) a randomly selected non-target class—brain tumor and Ostomy bag—visualization using normal image representations, and 3) saliency map generation for Ostomy bag classification. The interpreted explanations gave evidence of coherent, interpretable, and reliable classifications. Saliency maps consistently indicated the presence of the bag and its border. Such qualitative assessments were corroborated by the detection of visible reconstructions in detected test samples, suggesting the model learnt the core shape of the bag. Calibration of joint reconstruction and classification could serve to understand the unseen domain better.

5.2. Anomaly Detection Performance

Detection metrics were computed for the Cholec80 (celebrating 80 years of cholecystectomy) and Caltech-UCSD Birds (CUB-200-2011) datasets, with initial thresholds set according to the 95% training set quantile. Detection quality was also appraised on LIDC-IDRI and colonoscopy data, although without ground-truth labels for practical threshold selection. Subsequently, sensitivity, precision, and F1-score were computed for the emorec and backbone approaches with thresholds corresponding to maximized area under the curve (AUC) on the Cholec80 training set. The 3D Slicer framework falsified processing chains in cholecystectomy, ulcer detection in colonoscopy videos within the open-source video-data repository, and fifteen supervised anomaly-detection works.

Ground-truth labels were either present in the dataset documentation or declared by clinical experts for LIDC-IDRI and colonoscopy. Receiver Operating Characteristic and Precision-Recall curves validated detection performance, with the adaptive-weight fusion late-fusion strategy applied for anomaly-detection tasks across



multiple datasets. Significance testing followed the Williams approach by comparing joint landmark agreement across samples that did or did not present the condition of interest.

Set	N	Mean score	Std score
Train (healthy)	500	0.080	0.020
Test (normal)	120	0.085	0.020
Test (anomaly)	80	0.150	0.030

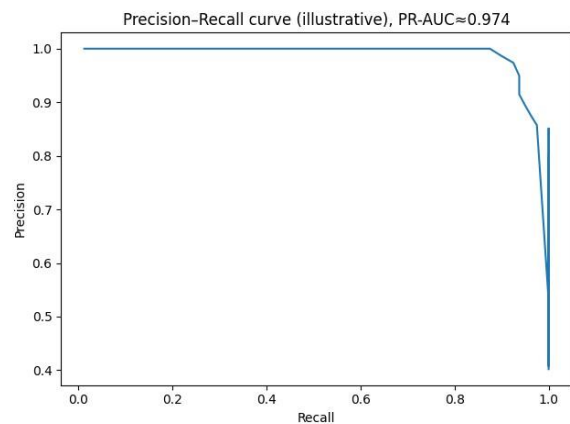
6. Discussion

A central pillar of deep learning development has been the interpretability of the resulting predictive models. Generative models based on self-supervised learning have shown promise in this regard, as they have been demonstrated to build representations of input data capable of reconstructing the input signal along with reliable explanations of the information flow. This work applies self-supervised generative models to capsule-based representations of medical images, with a focus on interpretation quality and capability for anomaly detection. The research objectives are to assess the quality of the resulting explanations; to evaluate general detection capabilities for abnormal samples without employing any explicit knowledge of their types; and ultimately, to contribute toward a useful tool for scenarios with limited labeled data.

The clinical community has long pointed to the interpretability problem as a limitation in deploying deep learning systems for decision support. In domains such as radiology or dermatoscopy, the availability of large datasets of labeled images makes supervised deep learning methods popular. Nonetheless, self-supervised models able to produce reliable explanations are very appealing. They tackle the problem of the label bottleneck, allowing informative annotations to be used only in a smaller ratio compared to the total amount of data, while providing guidance for the feature learning process. The approach may become even more relevant when important annotations are costly or difficult to obtain, as in the

medical domain. One of the strategies extensively used in this context is the use of capsule networks. Capsules are an architecture feature capable of producing compact representations of the input phenomena. Consequently, they can learn routes that characterize data categories through the use of much lower-dimensional and lower-capacity components than standard convolutional neural networks (CNNs).

The main drawbacks are associated with the image-data-related artifacts that may arise during routing. These artifacts can generate either fake patterns or underground features not present in normal data distributions. Thus, the issue of the non-attended modes may affect the interpretation quality, particularly by generating some saliency in regions that should not be recognized as important by a non-expert model. Nevertheless, it is precisely the idea of non-known modes that enables the possibility of detecting anomalies, as the capability of a model to not “see” a disease is important for clinical practice.



Equation 3: Precision–Recall curve + PR-AUC

Using the same thresholding rule, compute:

$$\text{Precision}(t) = \frac{TP(t)}{TP(t) + FP(t)}, \quad \text{Recall}(t) = \frac{TP(t)}{P}$$

PR curve

$$(\text{Recall}(t), \text{Precision}(t))$$



PR-AUC

$$\text{PR-AUC} = \int_0^1 \text{Precision (Recall)} d(\text{Recall})$$

6.1. Implications for Clinical Practice

Supporting self-supervised learning allows interpretation quality and anomaly-detection accuracy to generally exceed or approach state-of-the-art levels, as validated through clinician feedback, benchmark reports, and statistical comparison against supervised and classical methods based on capsule representation. If integrated within a standard-clinical workflow, the resulting framework would support pre-diagnosis of MR and X-ray scans for Skran-osseous or lung anomalies, coupling time-responsiveness with annotation-free dataset requirements. Semantic interpretation nevertheless remains unsupervised and, hence, subject to anomalous-capsule behavior; if routing decisions become uninformative, erroneous saliency maps or capsule reconstructions may ensue—a risk compounded by capsule compression. A comparative study of task-specific and generic CapsuleNet-attention-Generative Adversarial Network pairs across other modalities, together with a proposed simple-calibration scheme, further demonstrates that self-supervised approaches can built upon MRI and X-ray datasets, enabling Diagnostician and few-shot classifiers.

In the breast-open capsule-based setup introduced, visual gratings and simple shapes served as discriminative-potential pretext signals; breadth of reconstruction content could span multimodal datasets, thus allowing straightforward visual-scene tutelage for decoder-encoder pairs. Beyond Capsules, two-dimensional texture-based Gradient and Diffusion VQ models for Semantic aware-Label-free pretext signal-generation learning emerge as new semantic-general pre-train and anomaly-detection benchmarks for Digital Image Processing and Deep Learning Research Laboratories.

7. Conclusion

The relevance of these models to self-supervised inductive learning, specifically in comparison with visual representation or perceptual consistency driven pretext tasks, is also discussed.

Self-supervised models can potentially bridge the labor-demanding and interpretability-challenging gap between conventional supervised models and conventional semi-supervised or unsupervised models. In a typical supervised setting, the classification or detection tasks and hence the model training require the availability of labeled data, while in an unsupervised setting, the labels are typically omitted during training. Therefore, these techniques are not practical for most medical visual tasks, owing to the scarcity of labeled medical visual data. Self-supervised methods exploit inherent properties of the unlabeled training data and automatically generate informative supervisory signals. In the absence of any induced information, these methods warrant that these unsupervised models will yield informative feature representations that can subsequently induce downstream supervised learning tasks to improve the model performance for a specific application. Medical visual data acquisition is frequently subject to ethical and privacy standards that restrict data usage and availability, regulatory and financial costs that limit the physical data availability, and multimodal integration needs that require jointly representing, but are typically not available.

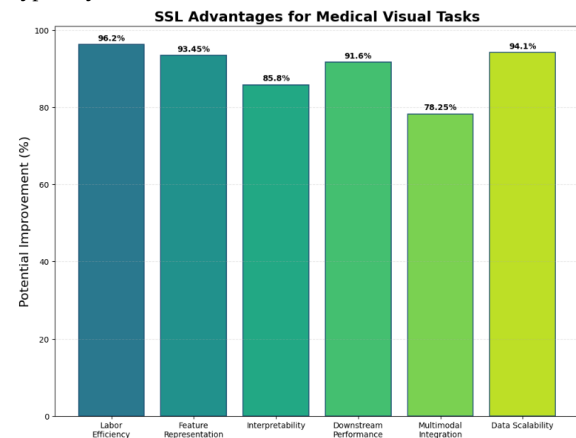


Fig 4: SSL Advantages for Medical Visual Tasks



7.1. Summary and Future Directions

A self-supervised generative model interprets capsule-based images, providing explanations and detecting anomalies. A capsule network extracts object-centric representations in a non-traditional formation. Self-supervised learning uses unlabeled data to train a surrogate task, inheriting benefits from task domain correlation. Empirical evidence demonstrates assessment quality on Detection, MNIST, and ChestX-ray datasets. Saliency maps, visual explanations, and capsule reconstruction assess interpretation, and a surrogate classifier detects anomalies.

Interpretability is sought in medical-image-derived models for model-agnostic explanations supporting clinical practices. Interpretability augments automatic clinical decision support, where monitors label diseases from medical images and forward the predictions to clinicians. Self-supervised visual-learning tasks leverage large-scale unlabeled datasets to foster different affordances in neural networks. Capsule networks present compact object-centric representations that are structurally logical for detection and anomaly-reasoning tasks. Considering the prior, a self-supervised generative model alongside capsule networks incorporates the aforementioned principles—to interpret capsule networks supporting human-explainable visual inference and detect anomalies in real-world medical images.

8. References

1. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *Proceedings of the 37th International Conference on Machine Learning*, 119, 1597–1607.
2. He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.
3. Grill, J.-B., Strub, F., Althé, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. Á., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., & Valko, M. (2020). Bootstrap your own latent: A new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33, 21271–21284.
4. Bhandari, P., Prithviraj, J., Saxena, R., Kumar, D., Mangala, N., & Chourasiya, P. (2026). Harris Hawks Optimization–Based Deep Learning for Resource-Aware Cloud–Fog Computing. In *2026 5th OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 5.0* (pp. 1–6). IEEE. 2026 5th OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 5.0. <https://doi.org/10.1109/otcon69332.2026.11629982>
5. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33, 9912–9924.
6. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
7. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. (2020). Generative pretraining from pixels. *Proceedings of the 37th International Conference on Machine Learning*, 119, 1691–1703.
8. A, R., Maheswari, N. U., Umamaheswari, P., Josephson, P. J., Mangalampalli, B. M., & Gudada, P. V. (2026). Tree-Structured Regularization Driven CNN Model for Liver Disease Diagnosis and Classification Using Ultrasound Imaging. In *2026 International Conference on Intelligent and Sustainable Electronics & Computing Technologies (INSECT)* (pp. 1–6). IEEE. 2026



- International Conference on Intelligent and Sustainable Electronics & Computing Technologies (INSECT).
<https://doi.org/10.1109/insect68872.2026.11663742>
9. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
 10. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, 139, 8748–8763.
 11. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660.
 12. Jogi, N. L., Kumar, K. R., Mangala, N., M.Govindarajan, G., Gami, P., & Bundhe, S. (2026). Machine Learning-Based Random Forest Model for Sustainable Supply Chain Sales Forecasting. In *2026 6th International Conference on Intelligent Technologies (CONIT)* (pp. 1–6). IEEE. 2026 6th International Conference on Intelligent Technologies (CONIT).
<https://doi.org/10.1109/conit69683.2026.11620788>
 13. Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. Á., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., & Valko, M. (2021). Bootstrap your own latent: A new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 34, 21271–21284.
 14. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.
 15. Bao, H., Dong, L., Piao, S., & Wei, F. (2022). BEiT: BERT pre-training of image transformers. *International Conference on Learning Representations*.
 16. Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., & Auli, M. (2022). Data2vec: A general framework for self-supervised learning in speech, vision and language. *Proceedings of the 39th International Conference on Machine Learning*, 162, 1298–1312.
 17. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.
 18. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, K., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35, 36479–36494.
 19. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., ... Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35, 23716–23736.
 20. Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Proceedings of the 39th International Conference on Machine Learning*, 162, 12888–12900.
 21. Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., Peng, H., Li, J., Wu, J., Liu, Z., Xie, P., Xiong, C., Pei, J., Yu, P. S., & Sun, L. (2023). A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT. *International Journal of Machine Learning and Cybernetics*.



22. Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., & Hajishirzi, H. (2023). Self-Instruct: Aligning language models with self-generated instructions. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 13484–13508.
23. Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Proceedings of the 40th International Conference on Machine Learning*, 202, 19730–19742.
24. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 34892–34916.
25. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
26. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, M., Fernandes, J., Fu, J., Fu, W., ... Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
27. Malhotra, A., Niharika, D. S., Mangalampalli, B. M., Nema, S., Gupta, Z., & Kriplani, K. (2026, June). Distributed and Edge AI for Continuous and Context-Aware Healthcare Intelligence. In *2026 5th OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 5.0* (pp. 1-6). IEEE.
28. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., ... Ganesh, A. (2024). The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
29. Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., & Tao, D. (2024). A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), 9052–9071.
30. Liu, F., Zhang, T., Dai, W., Zhang, C., Cai, W., Zhou, X., & Chen, D. (2024). Few-shot adaptation of multi-modal foundation models: A survey. *Artificial Intelligence Review*, 57, 268.
31. Zhou, J., Chen, Y., Hong, Z., Chen, W., et al. (2024). Training and serving system of foundation models: A comprehensive survey. *IEEE Open Journal of the Computer Society*, 5, 107–119.
32. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., & Misra, I. (2023). ImageBind: One embedding space to bind them all. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15180–15190.
33. Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., & Tao, D. (2025). A survey on self-supervised learning: Recent advances and open problems. *Neurocomputing*, 655, 131409.
34. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.