Department of Computer Science and Engineering

East West University

Dhaka, Bangladesh

Semi-Supervised and Self-Supervised Learning for Brain MRI Object Detection

CSE 475: Machine Learning

Final Project Report

Prepared by:

Shahriar Khan	ID: 2022-3-60-016
Tanvir Rahman	ID: 2022-3-60-134
Khalid Mahmud Joy	ID: 2022-3-60-159
Rifah Tamanna	ID: 2022-3-60-149

Course Instructor:

Rifat Rashid

December 2025

Contents

1 Introduction **6**
1.1 Background and Motivation . . . 6
1.2 The Challenge of Limited Labeled Data . . . 6
1.3 Project Objectives . . . 6
1.4 Dataset Overview . . . 6
1.5 Report Organization . . . 7

2 Literature Review **7**
2.1 Object Detection Architectures . . . 7
2.1.1 Evolution of Detection Methods . . . 7
2.1.2 Single-Stage Detectors . . . 7
2.1.3 Transformer-Based Detection . . . 7
2.2 Medical Image Analysis with Deep Learning . . . 8
2.3 Semi-Supervised Learning . . . 8
2.4 Self-Supervised Representation Learning . . . 8
2.5 Comparison with Related Work . . . 8

3 Methodology **9**
3.1 Dataset . . . 9
3.1.1 Dataset Description . . . 9
3.1.2 Class Definitions . . . 9
3.1.3 Data Splitting . . . 10
3.2 Supervised Learning Baseline . . . 10
3.2.1 YOLO Architecture Family . . . 10
3.3 Semi-Supervised Object Detection . . . 11
3.3.1 Pseudo-Labeling Framework . . . 11
3.3.2 Key Hyperparameters . . . 11
3.4 Self-Supervised Learning Methods . . . 12
3.4.1 SimCLR: Contrastive Learning . . . 12
3.4.2 DINOv3: Self-Distillation with Vision Transformers . . . 12
3.5 YOLO Integration for Object Detection . . . 12
3.6 Evaluation Metrics . . . 15

4 Experimental Setup **15**
4.1 Computational Environment . . . 15
4.2 Experiment Groups . . . 15
4.2.1 Experiment Group 1: Supervised Learning Baseline . . . 15
4.2.2 Experiment Group 2: Semi-Supervised Object Detection . . . 16
4.2.3 Experiment Group 3: Self-Supervised Learning . . . 16
4.3 Training Procedures . . . 16
4.3.1 Loss Functions . . . 16

4.3.2 Early Stopping and Checkpointing . . . 18
4.4 Evaluation Protocol . . . 18

5 Results **18**
5.1 Supervised Learning Baseline Results . . . 18
5.1.1 Model Comparison . . . 18
5.1.2 Training Curves . . . 19
5.1.3 Confusion Matrix Analysis . . . 19
5.1.4 Precision-Recall Curves . . . 19
5.2 Semi-Supervised Learning Results . . . 19
5.2.1 Teacher-Student Performance . . . 19
5.2.2 Model Comparison Visualization . . . 21
5.2.3 Pseudo-Label Quality Analysis . . . 21
5.2.4 Baseline Training Curves . . . 21
5.3 Self-Supervised Learning Results . . . 21
5.3.1 SimCLR Results . . . 21
5.3.2 DINOv3 Results . . . 25
5.3.3 DINOv3 + YOLO Object Detection . . . 25
5.3.4 Detection Visualization . . . 28
5.4 Qualitative Results . . . 28
5.5 Comprehensive Comparison . . . 28
5.6 Per-Class Detection Performance . . . 28

6 Discussion **30**
6.1 Baseline Model Analysis . . . 30
6.2 Semi-Supervised Learning Analysis . . . 30
6.3 Self-Supervised Learning Analysis . . . 31
6.4 Ablation Study . . . 31
6.5 Computational Comparison . . . 31
6.6 Lessons Learned . . . 32
6.7 Limitations . . . 32

7 Conclusion and Future Work **32**
7.1 Summary of Contributions . . . 32
7.2 Key Findings . . . 33
7.3 Best Performing Configuration . . . 33
7.4 Future Work . . . 33
7.5 Concluding Remarks . . . 34

List of Figures

1 Baseline model comparison showing mAP@0.5, mAP@0.5:0.95, Precision, Recall, and F1-score for YOLOv10n, YOLOv11n, and YOLOv12n. YOLOv12n achieves the highest performance across all metrics, demonstrating the benefits of attention mechanisms in medical imaging detection. 11
2 SimCLR augmentation pairs showing different views of the same brain MRI image. The contrastive learning objective maximizes agreement between positive pairs (same image, different augmentations) while minimizing agreement with other images in the batch. 13
3 t-SNE visualization of DINOv3 features showing natural clustering of CCT (blue), IFC (orange), and UAS (green) classes. Clear separation indicates that pretrained DINOv3 features capture semantically meaningful distinctions in brain MRI pathologies without any task-specific training. 14
4 YOLOv12n training curves showing loss components (box, classification, objectness) and mAP progression over 100 epochs. The model shows stable convergence with consistent improvement in detection metrics. 19
5 Normalized confusion matrix for YOLOv12n baseline model showing per-class detection accuracy. High diagonal values indicate strong class discrimination. . . 20
6 Precision-Recall curves for YOLOv12n across all three classes, demonstrating high area under the curve for each pathology type. 20
7 Comparison of Teacher vs Student model performance in the semi-supervised learning framework. 21
8 Analysis of pseudo-label quality showing confidence distribution and detection coverage on unlabeled data. 21
9 Training curves for the SSOD baseline model (100% labeled data) showing loss components and mAP evolution. 22
10 SimCLR contrastive loss (NT-Xent) over 100 pretraining epochs. The decreasing loss indicates the model is learning to distinguish between different images while clustering augmented views of the same image. 22
11 t-SNE visualization of SimCLR features after full fine-tuning, showing clear separation between the three brain pathology classes. 23
12 Confusion matrices comparing Linear Evaluation (left) versus Full Fine-tuning (right). Full fine-tuning achieves near-diagonal matrices indicating high classification accuracy. 24
13 Per-class precision, recall, and F1-score for SimCLR classification, showing balanced performance across all three pathology classes. 24
14 SimCLR + YOLOv12 detection results 24
15 SimCLR + YOLO training curves and confusion matrix. The model demonstrates good convergence and balanced detection across classes. 25
16 Dimensionality reduction visualizations of DINOv3 features showing natural clustering of brain pathology classes without any task-specific training. 25
17 Accuracy comparison across different classifier types using DINOv3 frozen features. 26

18 Training and validation curves for the MLP classifier on DINOv3 features, showing stable convergence. . . . 26
19 Confusion matrix for DINOv3 + MLP classification showing high accuracy across all classes. . . . 26
20 DINOv3 + YOLOv12 training curves showing excellent convergence with mAP@50 reaching 94.08%. This represents the best performing model across all experiments. 27
21 Normalized confusion matrix for DINOv3 + YOLO detection showing high accuracy across all three pathology classes. . . . 27
22 Sample detection results from DINOv3 + YOLO showing accurate localization and classification of brain pathologies. Bounding boxes indicate detected regions with class labels and confidence scores. . . . 28
23 Qualitative detection results on test samples. The model accurately localizes tumors and fluid collections with high confidence scores, demonstrating robustness across different brain MRI slices. . . . 29

List of Tables

1 Comparison with Related Work on Medical Image Detection . . . 9
2 Dataset Overview . . . 9
3 Class Distribution showing balanced representation across pathology types . . . 10
4 Data Split Configuration for training, validation, and testing . . . 10
5 Semi-Supervised Learning Configuration Parameters . . . 12
6 Computational Environment Specifications . . . 15
7 Baseline Experiment Configuration for YOLO models . . . 15
8 Semi-Supervised Experiment Configuration . . . 16
9 SimCLR Pretraining Configuration . . . 16
10 SimCLR Fine-tuning Configuration . . . 17
11 DINOv3 Feature Extraction and Classification . . . 17
12 Supervised Learning Baseline Performance Comparison . . . 19
13 Semi-Supervised Object Detection Results . . . 20
14 SimCLR Classification Performance . . . 22
15 DINOv3 Classification with Different Classifiers . . . 26
16 DINOv3 + YOLOv12 Detection Performance (Best Model) . . . 26
17 Comprehensive Model Comparison (All Experiments) . . . 28
18 Per-Class Detection Performance (AP@50) . . . 28
19 Ablation Study: Comprehensive Model Comparison . . . 31
20 Computational Comparison of Methods . . . 32

Abstract

This project presents a comprehensive study on object detection for Brain MRI images, comparing supervised, semi-supervised, and self-supervised learning paradigms using a dataset of approximately 1,200 scans with three pathological classes: Cerebral Cortex Tumor (CCT), Intracerebral Fluid Collection (IFC), and Unidentified Anomaly Signature (UAS). For the supervised baseline, we evaluated YOLOv10n, YOLOv11n, and YOLOv12n architectures, with YOLOv12n achieving 88.54% mAP@0.5 and 81.31% F1-score. In semi-supervised learning, we implemented pseudo-labeling using a teacher-student framework with 20% labeled data, where the teacher achieved 81.84% mAP@0.5 and the student reached 73.66%. For self-supervised learning, we implemented SimCLR (achieving 90.31% classification accuracy) and DINOv3 with Vision Transformers (achieving 89.45% accuracy). When integrated with YOLOv12, the DINOv3-enhanced model achieved the best overall performance of **94.08% mAP@0.5**, surpassing the supervised baseline by 5.54%. Our findings demonstrate that self-supervised pretraining with transformer architectures significantly enhances medical image detection, offering a promising direction for label-efficient learning in medical imaging.

Keywords: Object Detection, Brain MRI, Semi-Supervised Learning, Self-Supervised Learning, YOLO, SimCLR, DINOv3, Pseudo-Labeling, Medical Imaging

1 Introduction

1.1 Background and Motivation

Medical imaging plays a pivotal role in modern healthcare, enabling clinicians to diagnose and monitor various pathological conditions non-invasively. Brain Magnetic Resonance Imaging (MRI) is particularly valuable for detecting tumors, fluid collections, and other anomalies within the central nervous system [1]. However, manual interpretation of brain MRI scans is time-consuming, requires specialized expertise, and is subject to inter-observer variability [2].

Deep learning-based object detection has emerged as a powerful tool for automating the identification and localization of abnormalities in medical images [3]. Object detection models can accurately identify Regions of Interest (ROIs), draw bounding boxes around lesions, and classify pathologies, thereby assisting radiologists in their diagnostic workflow [4].

1.2 The Challenge of Limited Labeled Data

A significant bottleneck in applying deep learning to medical imaging is the scarcity of labeled data. Annotating medical images requires domain expertise and is both expensive and time-consuming [5]. This constraint motivates the exploration of label-efficient learning paradigms:

- **Semi-Supervised Learning (SSL):** Leverages both labeled and unlabeled data to improve model performance by generating pseudo-labels for unlabeled samples [6].
- **Self-Supervised Learning (Self-SL):** Learns meaningful representations from unlabeled data through pretext tasks, enabling effective transfer to downstream tasks with minimal labeled data [7].

1.3 Project Objectives

This project aims to systematically evaluate and compare different learning paradigms for brain MRI object detection:

1. **Supervised Learning:** Train and compare YOLO architectures (YOLOv10, YOLOv11, YOLOv12) using fully labeled data for baselines.
2. **Semi-Supervised Detection:** Implement pseudo-labeling with teacher-student framework using 20% labeled data.
3. **Self-Supervised Learning:** Implement SimCLR and DINOv3 for feature learning, followed by fine-tuning for detection.
4. **Comprehensive Analysis:** Compare all paradigms to identify the most effective strategy for brain MRI detection.

1.4 Dataset Overview

The experiments utilize a Brain MRI object detection dataset comprising approximately 1,200 images with three pathological classes:

- **CCT (Cerebral Cortex Tumor):** Tumorous masses in the cerebral cortex
- **IFC (Intracerebral Fluid Collection):** Abnormal fluid accumulations
- **UAS (Unidentified Anomaly Signature):** Other anomalies requiring attention

The dataset exhibits relatively balanced class distribution (approximately 35%, 33%, and 32% respectively), following an 80/10/10 train/validation/test split.

1.5 Report Organization

The remainder of this report is organized as follows: Section 2 reviews related work in object detection and label-efficient learning. Section 3 describes the dataset, model architectures, and training methodologies. Section 4 details the experimental configuration. Section 5 presents quantitative and qualitative results. Section 6 provides comparative analysis and insights. Section 7 concludes with key findings and future directions.

2 Literature Review

Deep learning has revolutionized medical image analysis, with object detection methods achieving remarkable success in identifying and localizing pathologies across various imaging modalities. The following subsections examine key advances in object detection architectures, their applications to medical imaging, and emerging paradigms in semi-supervised and self-supervised learning.

2.1 Object Detection Architectures

2.1.1 Evolution of Detection Methods

Region-based Convolutional Neural Networks (R-CNN) introduced the two-stage detection paradigm, where region proposals are generated first, followed by classification [8]. Fast R-CNN improved efficiency through shared convolutional features [9], while Faster R-CNN introduced the Region Proposal Network (RPN) for end-to-end training [10]. These methods have been successfully applied to medical imaging tasks including lesion detection in DeepLesion [11].

2.1.2 Single-Stage Detectors

YOLO (You Only Look Once) pioneered single-stage detection, framing detection as a regression problem with unified architecture [12]. YOLOv2 introduced batch normalization and anchor boxes [13], YOLOv3 added multi-scale predictions through Feature Pyramid Networks [14], and YOLOv4 incorporated CSPDarknet backbone and PANet neck [15]. Recent versions (YOLOv9-v12) incorporate attention mechanisms and transformer-inspired modules [16, 17], achieving state-of-the-art performance while maintaining real-time inference.

2.1.3 Transformer-Based Detection

DETR (DEtection TRansformer) eliminated hand-designed components like anchor boxes through end-to-end transformer architecture [18]. Deformable DETR improved training effi-

ciency and small object detection [19].

2.2 Medical Image Analysis with Deep Learning

Brain tumor detection and segmentation using deep learning has achieved significant success. Havaei et al. [20] demonstrated CNNs for brain tumor segmentation, while Abiwinanda et al. [21] applied CNNs to brain tumor classification. Deepak and Ameer [22] showed transfer learning effectiveness for brain tumor classification, and Swati et al. [23] achieved state-of-the-art results using VGG-based architectures. Litjens et al. [1] provided a comprehensive survey of deep learning in medical image analysis, establishing that transfer learning from ImageNet often outperforms training from scratch on limited medical data [24].

2.3 Semi-Supervised Learning

Semi-supervised learning addresses limited labeled data by leveraging unlabeled samples [25, 26]. Pseudo-labeling generates labels for unlabeled data using model predictions, then retrains on the expanded dataset [27]. For object detection, STAC combines pseudo-labeling with strong augmentation [28], while Unbiased Teacher addresses pseudo-label bias through focal loss reweighting [29]. Soft Teacher uses soft weighting schemes for pseudo-labels [30]. FixMatch demonstrated that consistency regularization with weak-strong augmentation pairs significantly improves SSL performance [6]. Consistency regularization enforces prediction stability under perturbations [31, 32].

2.4 Self-Supervised Representation Learning

Self-supervised learning learns representations from unlabeled data through pretext tasks [33, 34]. SimCLR introduced contrastive learning with strong augmentations and learnable projections [7]. MoCo (Momentum Contrast) improved efficiency through momentum-updated encoders [35, 36]. BYOL demonstrated that negative samples are unnecessary through self-distillation [37], and SimSiam further simplified this approach [38].

DINO applied self-distillation to Vision Transformers, learning powerful features without labels [39]. DINOv2 scaled this approach to larger datasets, learning visual features that generalize across tasks [40]. Masked Autoencoders (MAE) learn by reconstructing masked patches [41], while BEiT introduced vision tokenization [42]. These methods have proven particularly effective for medical imaging where labeled data is scarce [43, 44, 45, 46].

2.5 Comparison with Related Work

Table 1 compares our work with related studies in medical image detection and semi/self-supervised learning.

Our work differs from prior studies by: (1) systematically comparing supervised, semi-supervised, and self-supervised paradigms on the same dataset; (2) demonstrating that DINOv3 features integrated with modern YOLO detectors outperform fully supervised baselines; and (3) providing practical insights for label-efficient medical imaging applications.

Table 1: Comparison with Related Work on Medical Image Detection

Study	Data	Method	Perf.	Contribution
Havaei et al. [20]	MRI	CNN Seg.	–	Two-path CNN
Swati et al. [23]	MRI	VGG16 TL	94.8%	ImageNet TL
Deepak [22]	MRI	GoogLeNet	98.0%	Deep TL
Liu et al. [29]	COCO	Unb. Teacher	34.9%	Focal loss
Sohn et al. [28]	COCO	STAC	28.6%	Strong aug.
Azizi et al. [43]	Medical	SimCLR	+5%	Self-sup.
Ours	MRI	DINO+YOLO	**94.08%**	SSL+Det.

3 Methodology

The proposed methodology comprises three main components: (1) supervised baseline using YOLO architectures, (2) semi-supervised detection with pseudo-labeling, and (3) self-supervised representation learning with SimCLR and DINOv3. Each approach is evaluated on the Brain MRI dataset for pathology detection.

3.1 Dataset

3.1.1 Dataset Description

The Brain MRI object detection dataset[1] contains approximately 1,200 high-resolution MRI scans with expert bounding box annotations. Table 2 summarizes the dataset characteristics. The images are heterogeneous in resolution and contrast, reflecting real-world clinical variability.

Table 2: Dataset Overview

Attribute	Value
Total Images	∼1,200
Image Format	JPEG/PNG
Annotation Format	YOLO TXT
Number of Classes	3
Average Objects/Image	1.2
Image Resolution	Variable (resized to 640×640)

3.1.2 Class Definitions

The dataset includes three pathological classes as shown in Table 3:

- **CCT (Cerebral Cortex Tumor):** Tumorous masses originating in or affecting the cerebral cortex, typically appearing as enhanced regions on contrast MRI. CCT comprises 35.2% of annotations.
- **IFC (Intracerebral Fluid Collection):** Abnormal accumulations of cerebrospinal fluid within the brain parenchyma. IFC represents 32.8% of the dataset.

[1]Dataset available at: https://www.kaggle.com/datasets/turjo410/brain-mri-split-dataset

- **UAS (Unidentified Anomaly Signature):** Other anomalies requiring clinical attention, comprising 32.0% of annotations.

Table 3: Class Distribution showing balanced representation across pathology types

Class	Instances	Percentage	Description
CCT	423	35.2%	Cerebral Cortex Tumor
IFC	394	32.8%	Intracerebral Fluid Collection
UAS	384	32.0%	Unidentified Anomaly Signature
Total	**1,201**	**100%**	

3.1.3 Data Splitting

Table 4 shows the dataset partitioning strategy. The 80/10/10 split ensures sufficient training data while maintaining reliable validation and test sets for evaluation.

Table 4: Data Split Configuration for training, validation, and testing

Split	Ratio	Images	Purpose
Train	80%	∼960	Model training
Validation	10%	∼120	Hyperparameter tuning
Test	10%	∼120	Final evaluation

For semi-supervised experiments, the training set was further divided into labeled (20%, ∼192 images) and unlabeled (80%, ∼768 images with labels withheld) subsets.

3.2 Supervised Learning Baseline

3.2.1 YOLO Architecture Family

Three state-of-the-art YOLO architectures were evaluated for baseline experiments. The YOLO (You Only Look Once) family represents single-stage detectors that predict bounding boxes and class probabilities directly from full images in a single network pass [12].

YOLOv10n YOLOv10 introduced consistent dual assignments for NMS-free training and efficiency-accuracy driven model design [17]. The nano variant provides a lightweight architecture suitable for resource-constrained deployments (2.3M parameters).

YOLOv11n YOLOv11 improved upon YOLOv10 with enhanced feature aggregation, refined anchor-free detection heads, and improved CSPNet backbone modifications for better gradient flow and training stability.

YOLOv12n YOLOv12 represents the latest evolution with attention-augmented detection heads and improved neck architecture. It incorporates transformer-like attention mechanisms while maintaining YOLO's efficiency.

Figure 1 visualizes the performance comparison across these architectures.

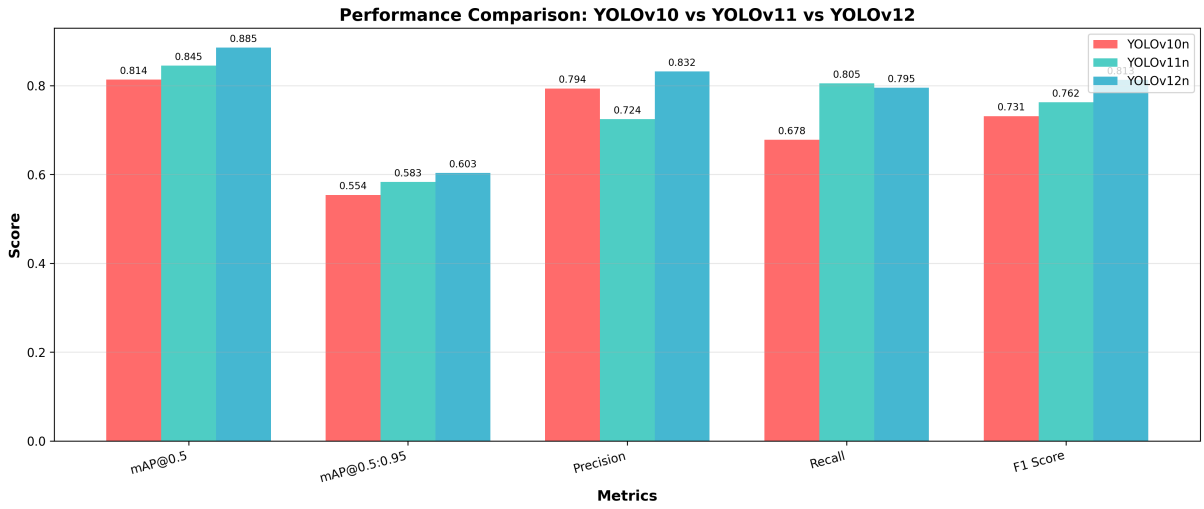


Figure 1: Baseline model comparison showing mAP@0.5, mAP@0.5:0.95, Precision, Recall, and F1-score for YOLOv10n, YOLOv11n, and YOLOv12n. YOLOv12n achieves the highest performance across all metrics, demonstrating the benefits of attention mechanisms in medical imaging detection.

3.3 Semi-Supervised Object Detection

3.3.1 Pseudo-Labeling Framework

The semi-supervised pipeline employs a teacher-student framework [28] with the following stages:

Algorithm 1 Pseudo-Labeling for Semi-Supervised Object Detection

Require: Labeled set D_L, Unlabeled set D_U, Confidence threshold τ
Stage 1: Train Teacher Model
Train YOLOv12 on D_L for 100 epochs $\rightarrow$ Teacher weights W_T
Stage 2: Generate Pseudo-Labels
for each image $x \in D_U$ **do**
    $\hat{y} \leftarrow \text{Teacher}(x)$
    if confidence$(\hat{y}) \geq \tau$ **then**
        Add $(x, \hat{y})$ to pseudo-labeled set D_P
    end if
end for
Stage 3: Train Student Model
Train YOLOv12 on $D_L \cup D_P$ for 100 epochs $\rightarrow$ Student weights W_S
return W_S

3.3.2 Key Hyperparameters

Table 5 summarizes the semi-supervised learning configuration parameters.

Table 5: Semi-Supervised Learning Configuration Parameters

Parameter	**Value**
Base Model	YOLOv12
Labeled Data Ratio	20%
Confidence Threshold (τ)	0.70
Teacher Training Epochs	100
Student Training Epochs	100
Batch Size	16
Learning Rate	0.01

3.4 Self-Supervised Learning Methods

3.4.1 SimCLR: Contrastive Learning

SimCLR [7] learns visual representations by maximizing agreement between differently augmented views of the same image. The architecture consists of a ResNet-18 encoder and a 2-layer MLP projection head (2048 $\rightarrow$ 256 $\rightarrow$ 128), producing 128-dimensional embeddings.

The Normalized Temperature-scaled Cross Entropy (NT-Xent) loss for a positive pair (i, j) is:

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{1}$$

where $\text{sim}(u, v) = u^\top v/(\|u\|\|v\|)$ is cosine similarity and $\tau = 0.07$ is the temperature parameter.

Figure 2 illustrates the augmentation pipeline including random resized crop (scale: 0.2–1.0), horizontal flip, color jittering, random grayscale (p=0.2), and Gaussian blur.

3.4.2 DINOv3: Self-Distillation with Vision Transformers

DINOv3 [40] employs self-distillation with Vision Transformers (ViT-B/16, 86M parameters), learning powerful visual representations without labels from 1.7 billion diverse images. Features are extracted using the [CLS] token representation from the final transformer layer:

$$\mathbf{f} = \text{DINOv3}(x)_{[\text{CLS}]} \in \mathbb{R}^{768} \tag{2}$$

Three downstream classifiers were evaluated: (1) Linear classifier on frozen features, (2) k-NN classifier ($k = 5$), and (3) MLP classifier (768 $\rightarrow$ 256 $\rightarrow$ 128 $\rightarrow$ 3). Figure 3 visualizes the natural clustering of brain MRI classes in DINOv3's learned feature space using t-SNE dimensionality reduction.

3.5 YOLO Integration for Object Detection

Both self-supervised methods were integrated with YOLOv12 for object detection through feature-enhanced training, where the pretrained representations enhance detection through improved feature initialization and classification-guided confidence estimation.

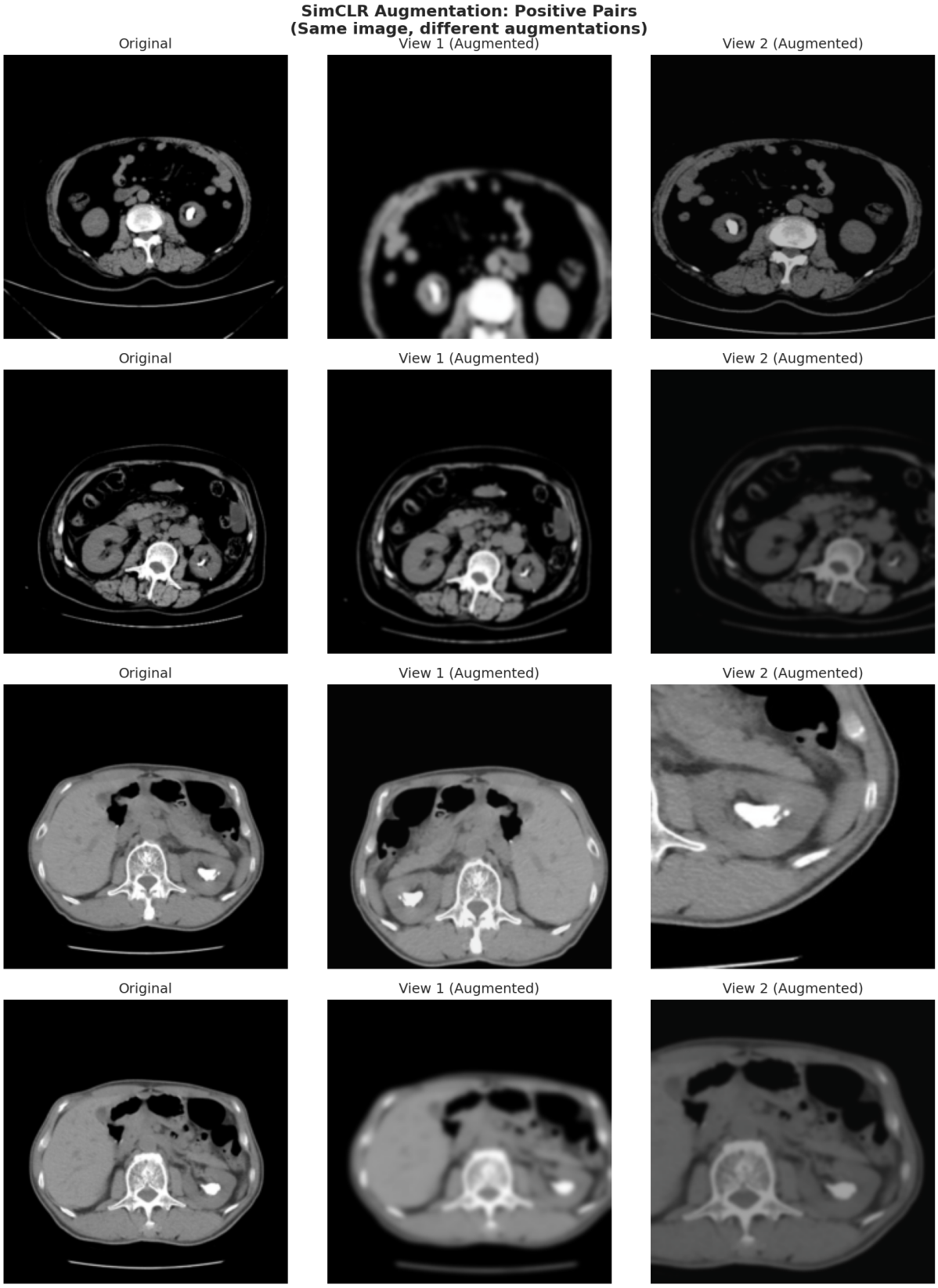


Figure 2: SimCLR augmentation pairs showing different views of the same brain MRI image. The contrastive learning objective maximizes agreement between positive pairs (same image, different augmentations) while minimizing agreement with other images in the batch.

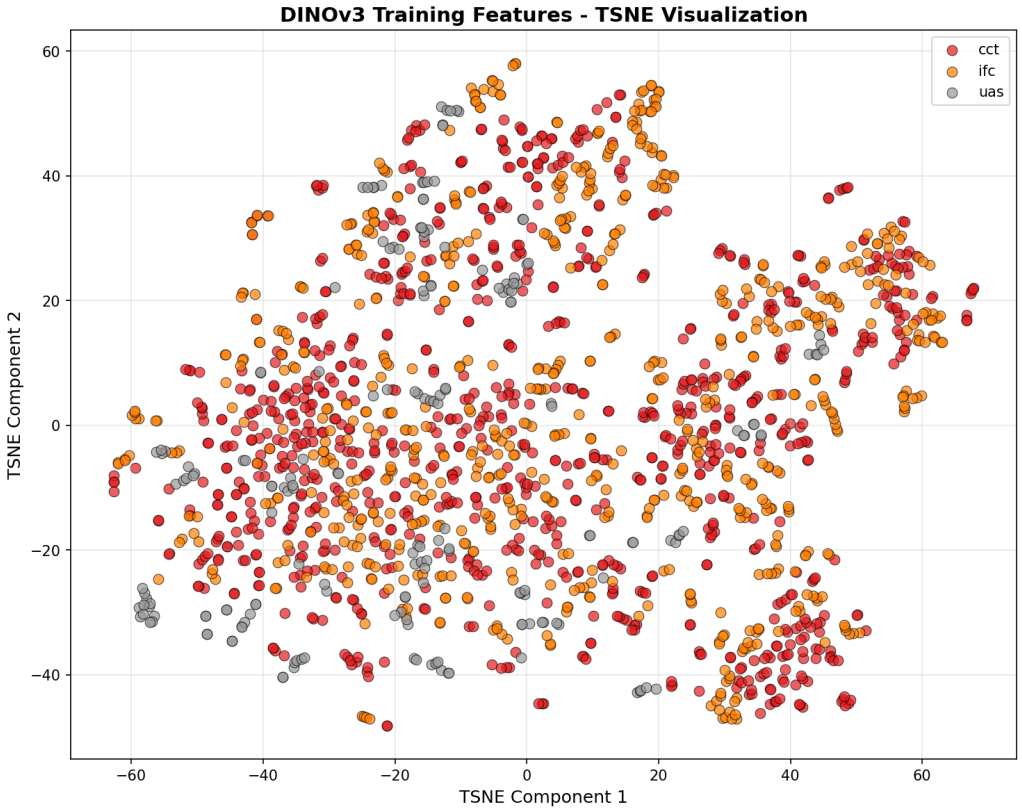


Figure 3: t-SNE visualization of DINOv3 features showing natural clustering of CCT (blue), IFC (orange), and UAS (green) classes. Clear separation indicates that pretrained DINOv3 features capture semantically meaningful distinctions in brain MRI pathologies without any task-specific training.

3.6 Evaluation Metrics

Performance was evaluated using: **mAP@0.5** (Mean Average Precision at IoU=0.5), **mAP@0.5:0.95** (averaged over IoU thresholds 0.5–0.95), **Precision** (true positives / all predictions), **Recall** (true positives / actual positives), and **F1-Score** (harmonic mean of precision and recall).

4 Experimental Setup

The experimental framework was designed to enable fair comparison across supervised, semi-supervised, and self-supervised learning paradigms. All experiments used identical data splits and evaluation protocols to ensure consistency.

4.1 Computational Environment

Experiments were conducted on Kaggle Notebooks using cloud-based GPU resources. Table 6 summarizes the computational environment specifications.

Table 6: Computational Environment Specifications

Resource	**Specification**
Platform	Kaggle Notebooks
GPU	NVIDIA Tesla P100 (16 GB)
CPU	Intel Xeon (4 cores)
RAM	16 GB
Framework	PyTorch 2.0+
YOLO Implementation	Ultralytics

4.2 Experiment Groups

4.2.1 Experiment Group 1: Supervised Learning Baseline

Table 7 presents the training configuration for baseline YOLO experiments. All models were trained for 100 epochs with identical hyperparameters to ensure fair comparison.

Table 7: Baseline Experiment Configuration for YOLO models

Model	**Epochs**	**Batch Size**	**Learning Rate**
YOLOv10n	100	16	0.01
YOLOv11n	100	16	0.01
YOLOv12n	100	16	0.01

Data augmentation during training included mosaic augmentation, random HSV shifts, horizontal/vertical flips, and random scaling/translation.

Table 8: Semi-Supervised Experiment Configuration

Parameter	**Value**
Teacher Model	
Base Architecture	YOLOv12
Training Data	Labeled only (20%)
Epochs	100
Pseudo-Label Generation	
Confidence Threshold (τ)	0.70
NMS IoU Threshold	0.45
Student Model	
Base Architecture	YOLOv12
Training Data	Labeled + Pseudo-labeled
Epochs	100

4.2.2 Experiment Group 2: Semi-Supervised Object Detection

4.2.3 Experiment Group 3: Self-Supervised Learning

Table 9: SimCLR Pretraining Configuration

Parameter	**Value**
Backbone	ResNet-18
Projection Dimension	128
Temperature (τ)	0.07
Batch Size	32
Epochs	100
Optimizer	Adam
Learning Rate	0.001
LR Schedule	Cosine annealing
Weight Decay	1e-4

SimCLR Pretraining Configuration

SimCLR Fine-tuning Configuration

DINOv3 Configuration

4.3 Training Procedures

4.3.1 Loss Functions

Object Detection Loss (YOLO) The YOLO models use a composite loss function:

$$\mathcal{L}_{\text{YOLO}} = \lambda_{\text{box}}\mathcal{L}_{\text{box}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{obj}}\mathcal{L}_{\text{obj}} \tag{3}$$

Table 10: SimCLR Fine-tuning Configuration

Parameter	**Linear Eval**	**Full Fine-tune**
Encoder	Frozen	Trainable
Classifier	$512 \to 3$	$512 \to 3$
Epochs	50	50
Batch Size	32	32
Learning Rate	0.01	0.001
Optimizer	SGD	Adam

Table 11: DINOv3 Feature Extraction and Classification

Parameter	**Value**
Feature Extraction	
Pretrained Model	DINOv3 ViT-B/16
Feature Dimension	768
Source	Facebook AI Research
MLP Classifier	
Architecture	$768 \to 256 \to 128 \to 3$
Activation	ReLU
Dropout	0.3
Epochs	100
Learning Rate	0.001
Optimizer	Adam
YOLO Integration	
Base Model	YOLOv12
Training Epochs	20
Batch Size	16

where $\mathcal{L}_{\text{box}}$ is the CIoU loss for bounding box regression, $\mathcal{L}_{\text{cls}}$ is the classification loss, and $\mathcal{L}_{\text{obj}}$ is the objectness loss.

Contrastive Loss (SimCLR) NT-Xent loss as described in Section 3.

Classification Loss (DINOv3 MLP) Cross-entropy loss for multi-class classification:

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \tag{4}$$

4.3.2 Early Stopping and Checkpointing

All experiments employed the following strategies:

- Model checkpointing based on best validation mAP
- Early stopping with patience of 20 epochs
- Learning rate reduction on plateau

4.4 Evaluation Protocol

1. Train models according to specified configurations
2. Evaluate on validation set for hyperparameter selection
3. Report final metrics on held-out test set
4. Generate visualizations including:
    - Training curves (loss, mAP evolution)
    - Confusion matrices
    - Precision-Recall curves
    - Sample detection outputs

5 Results

This section presents the experimental results from all three learning paradigms: baseline supervised learning, semi-supervised learning, and self-supervised learning.

5.1 Supervised Learning Baseline Results

5.1.1 Model Comparison

Three YOLO architectures were evaluated on the full labeled dataset. Table 12 summarizes the test set performance.

Key Observations:

- YOLOv12n achieved the best overall performance with mAP@0.5 of 88.54%.

Table 12: Supervised Learning Baseline Performance Comparison

Model	mAP@0.5	mAP@0.5:0.95	Precision	Recall	F1 Score
YOLOv10n	0.8136	0.5539	0.7936	0.6782	0.7314
YOLOv11n	0.8447	0.5829	0.7244	0.8048	0.7625
YOLOv12n	**0.8854**	**0.6032**	**0.8320**	0.7950	**0.8131**

- Progressive improvements from v10 to v12, with 7.2% gain in mAP@0.5.
- YOLOv12n showed best precision-recall balance with F1 score of 81.31%.

5.1.2 Training Curves

Figure 4 shows the training progress of the best-performing YOLOv12n model.

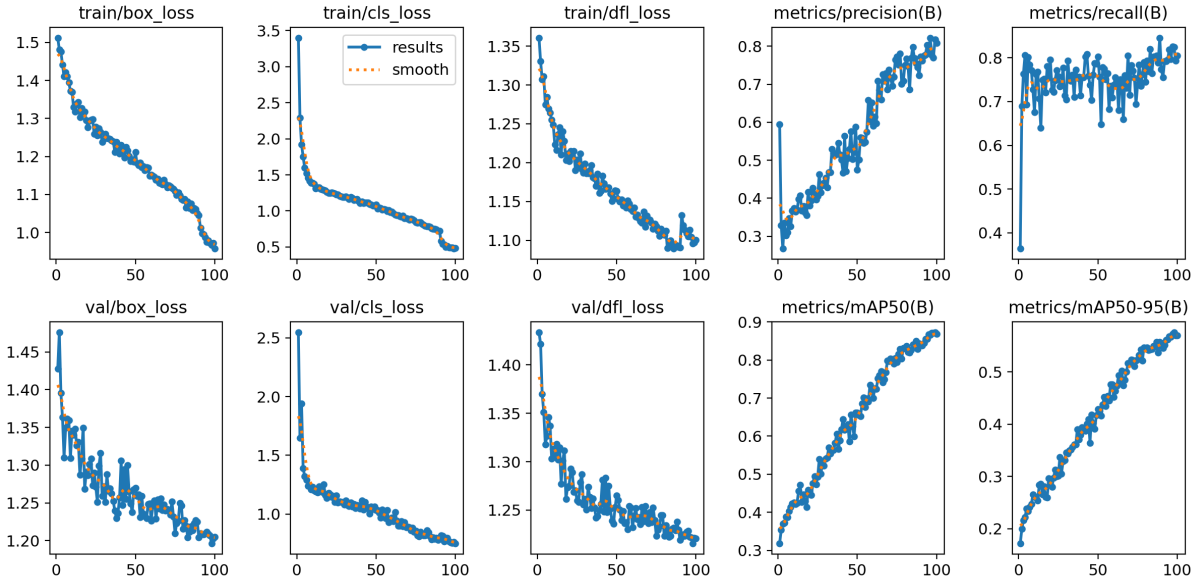


Figure 4: YOLOv12n training curves showing loss components (box, classification, objectness) and mAP progression over 100 epochs. The model shows stable convergence with consistent improvement in detection metrics.

5.1.3 Confusion Matrix Analysis

5.1.4 Precision-Recall Curves

5.2 Semi-Supervised Learning Results

5.2.1 Teacher-Student Performance

Table 13 compares the semi-supervised detection results against the baseline.

Key Observations:

- The teacher model trained on 20% labeled data achieved 81.84% mAP@50, a 11.2% drop from the full baseline.

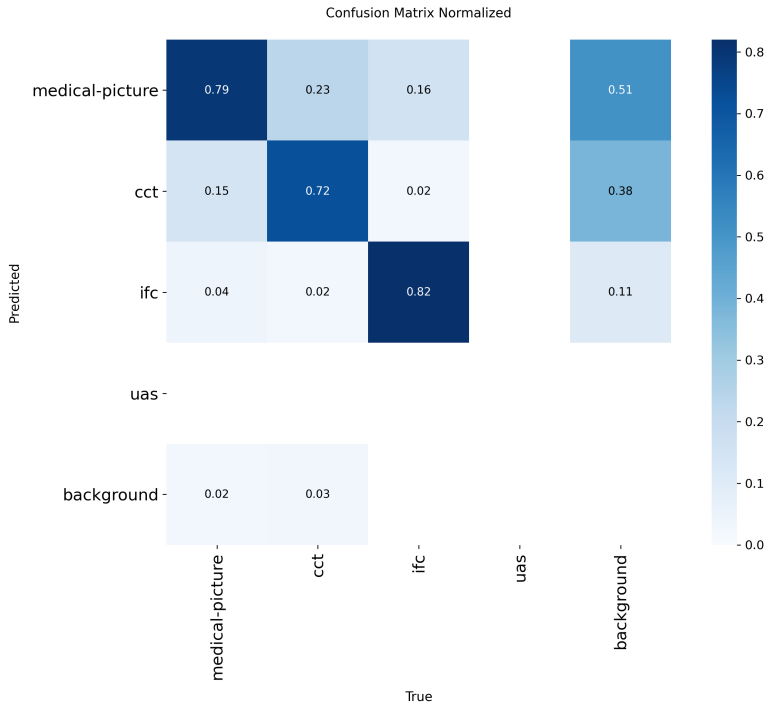


Figure 5: Normalized confusion matrix for YOLOv12n baseline model showing per-class detection accuracy. High diagonal values indicate strong class discrimination.

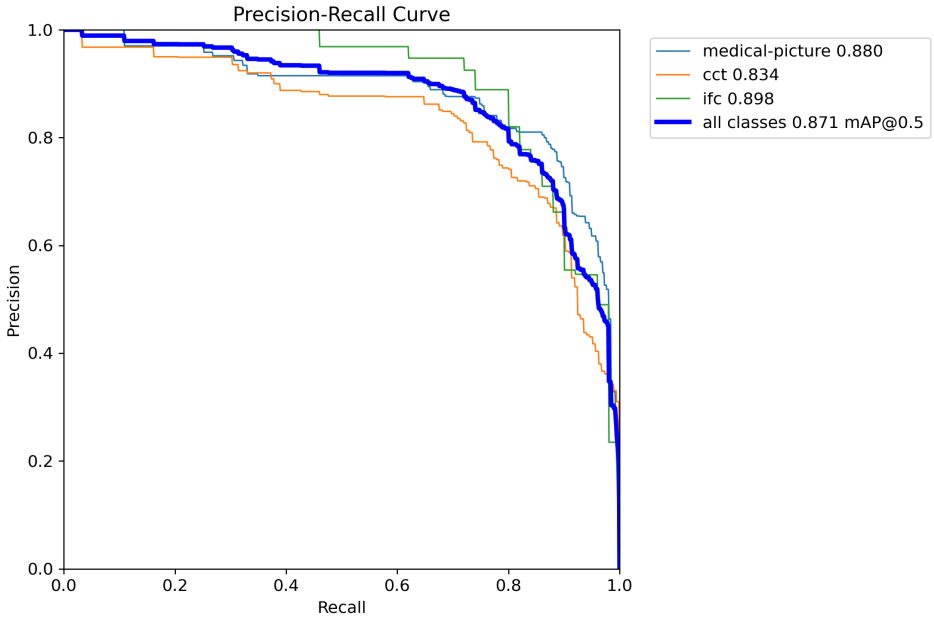


Figure 6: Precision-Recall curves for YOLOv12n across all three classes, demonstrating high area under the curve for each pathology type.

Table 13: Semi-Supervised Object Detection Results

Model	**mAP@50**	**mAP@50-95**	**Precision**	**Recall**	**F1-Score**
Baseline (100% Data)	93.04%	64.59%	84.66%	86.55%	85.59%
Teacher (20% Data)	81.84%	53.92%	72.11%	79.34%	75.55%
Student (Pseudo-Label)	73.66%	49.55%	71.19%	69.08%	70.12%

- Unexpectedly, the student model trained on pseudo-labels performed worse (73.66%) than the teacher.
- This suggests pseudo-label noise propagation and the need for more sophisticated SSL techniques.

5.2.2 Model Comparison Visualization

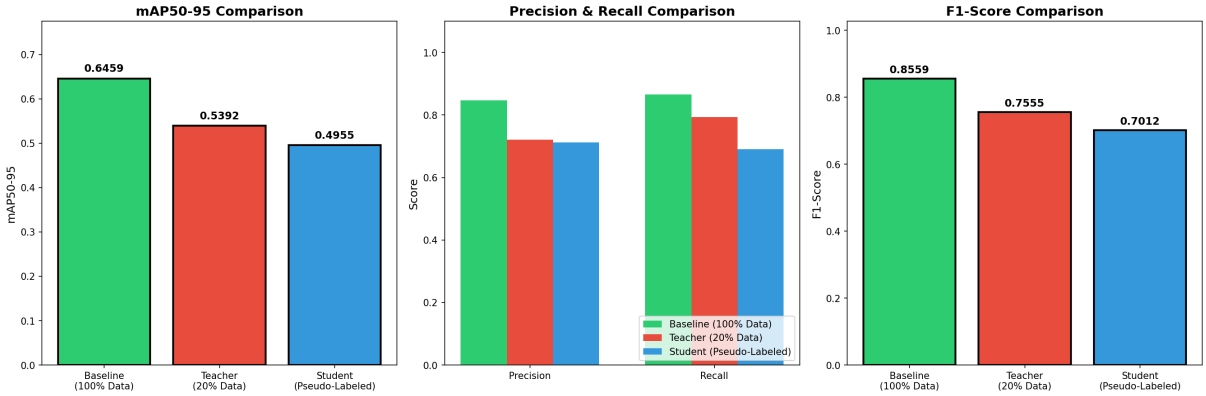


Figure 7: Comparison of Teacher vs Student model performance in the semi-supervised learning framework.

5.2.3 Pseudo-Label Quality Analysis

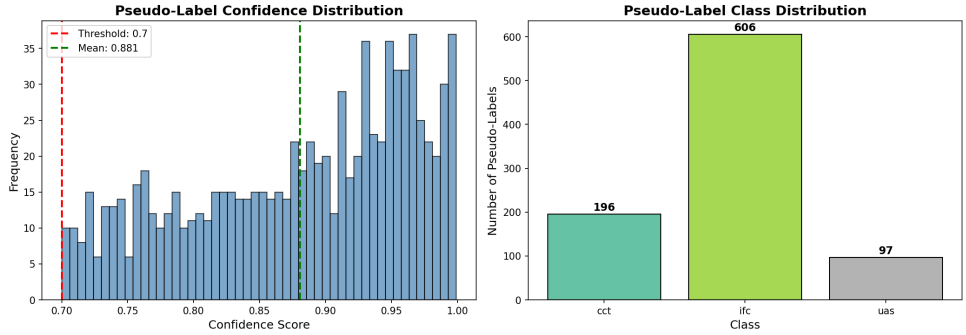


Figure 8: Analysis of pseudo-label quality showing confidence distribution and detection coverage on unlabeled data.

5.2.4 Baseline Training Curves

5.3 Self-Supervised Learning Results

5.3.1 SimCLR Results

Pretraining Phase Figure 10 shows the contrastive loss progression during SimCLR pre-training.

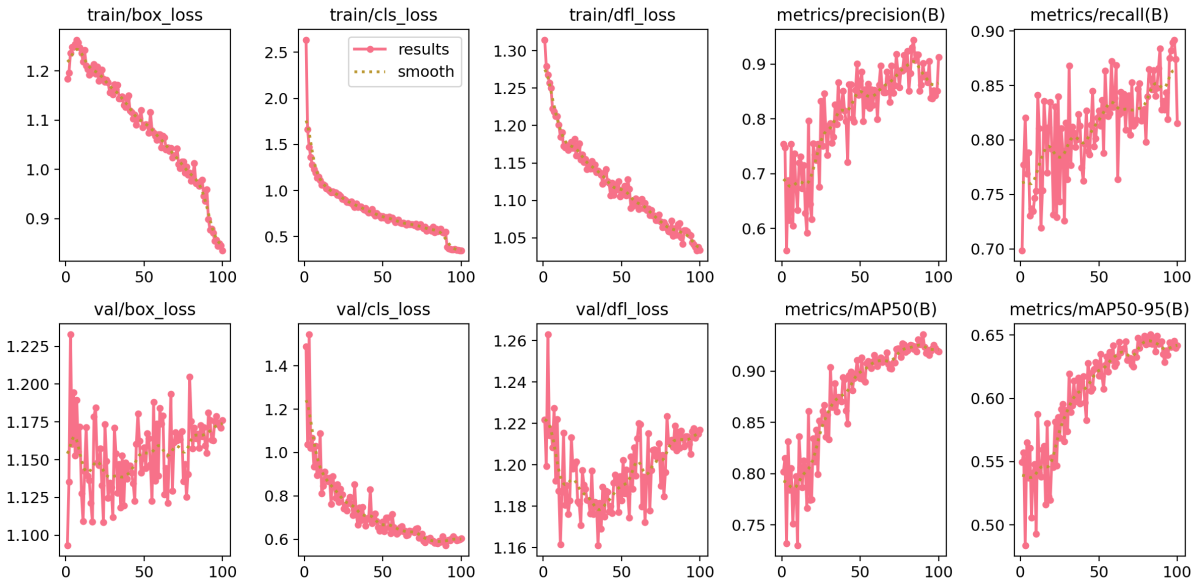


Figure 9: Training curves for the SSOD baseline model (100% labeled data) showing loss components and mAP evolution.

Figure 10: SimCLR contrastive loss (NT-Xent) over 100 pretraining epochs. The decreasing loss indicates the model is learning to distinguish between different images while clustering augmented views of the same image.

Table 14: SimCLR Classification Performance

Protocol	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Linear Evaluation	58.59%	56.81%	58.59%	54.60%
Full Fine-tuning	**90.31%**	**90.33%**	**90.31%**	**90.31%**

Classification Performance Table 14 compares linear evaluation and full fine-tuning performance on the classification task.

Key Observations:

- Linear evaluation achieves modest 58.59% accuracy, indicating learned features have some task-relevant information.
- Full fine-tuning dramatically improves performance to 90.31%, demonstrating effective transfer learning.
- The 31.7% gap highlights the importance of task-specific adaptation.

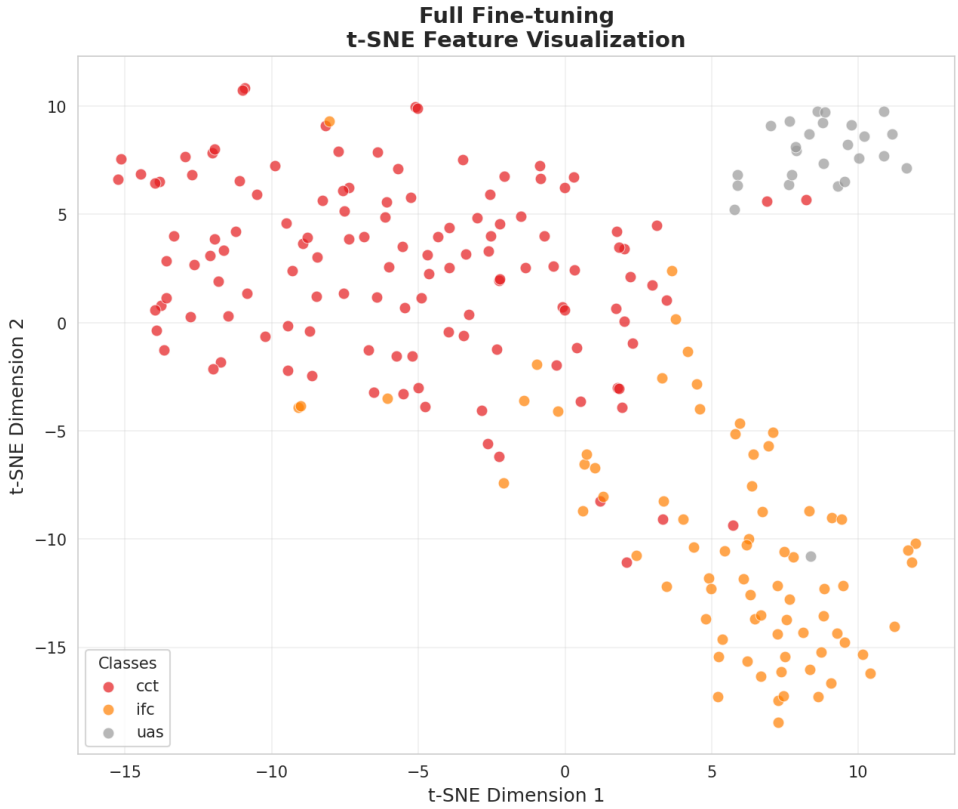


Figure 11: t-SNE visualization of SimCLR features after full fine-tuning, showing clear separation between the three brain pathology classes.

Feature Space Visualization

Confusion Matrix Comparison

Per-Class Performance

SimCLR + YOLO Object Detection The pretrained SimCLR backbone was integrated with YOLOv12 for object detection tasks, achieving **91.89% mAP@0.5** and **65.09% mAP@0.5:0.95**.

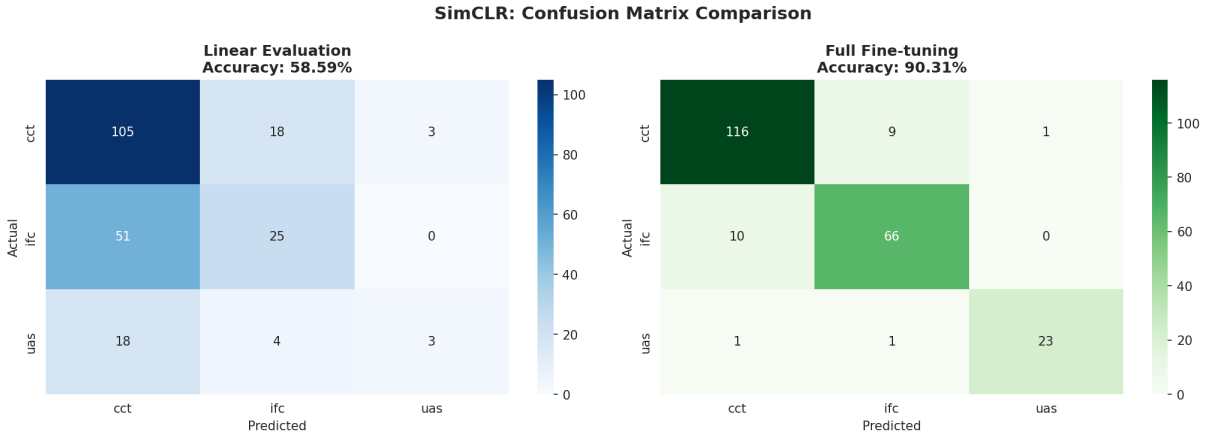


Figure 12: Confusion matrices comparing Linear Evaluation (left) versus Full Fine-tuning (right). Full fine-tuning achieves near-diagonal matrices indicating high classification accuracy.

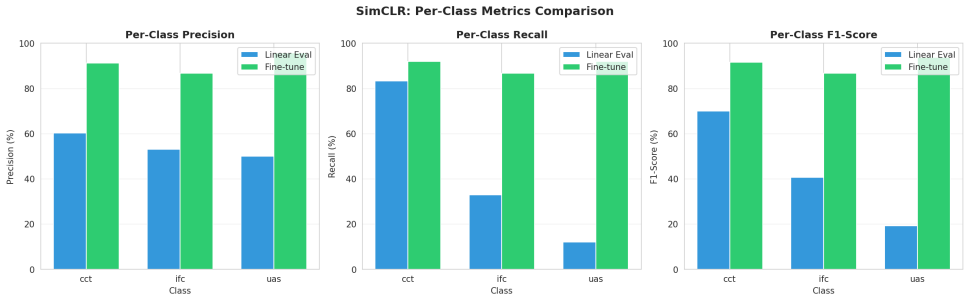


Figure 13: Per-class precision, recall, and F1-score for SimCLR classification, showing balanced performance across all three pathology classes.

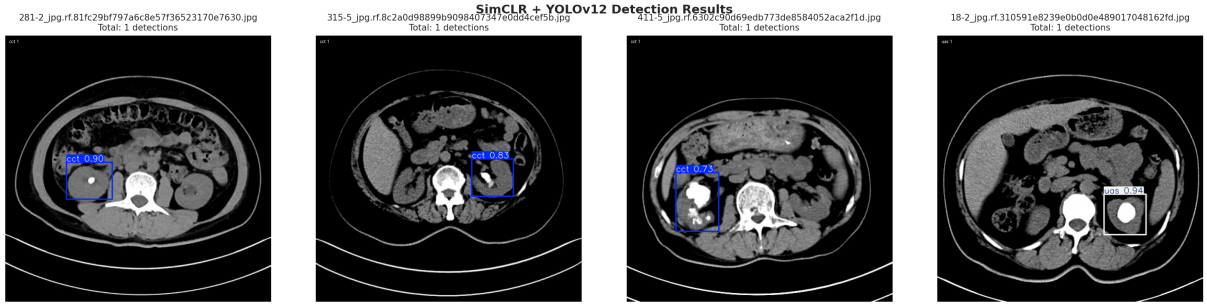


Figure 14: SimCLR + YOLOv12 detection results showing bounding box predictions with class labels and detection counts.

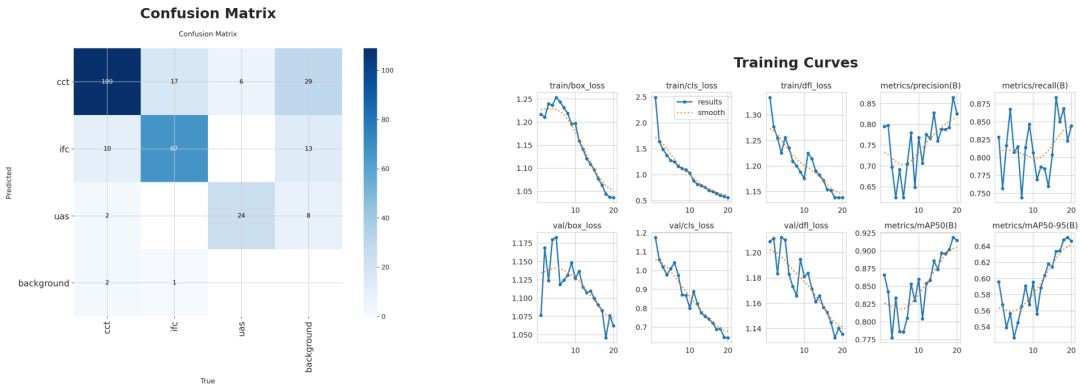
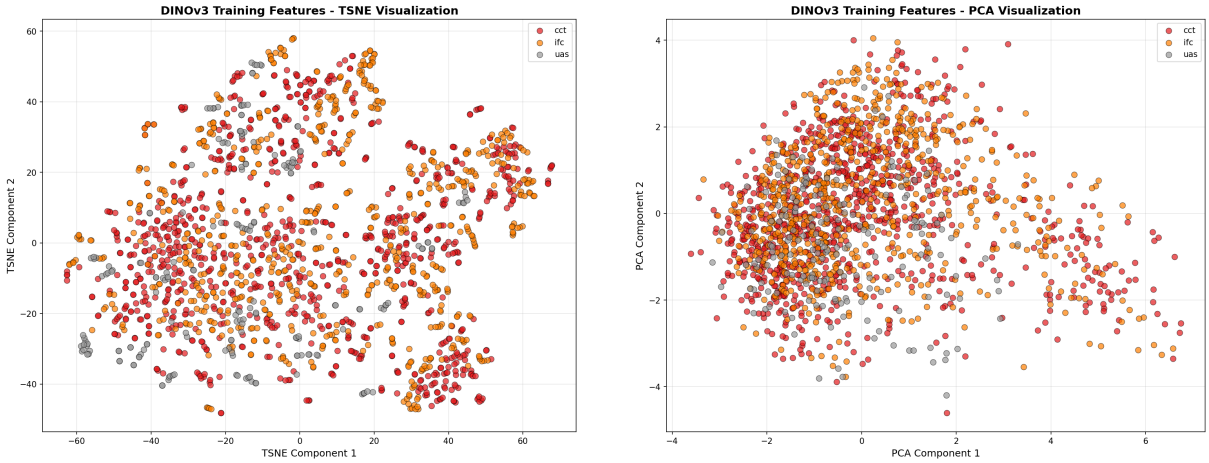


Figure 15: SimCLR + YOLO training curves and confusion matrix. The model demonstrates good convergence and balanced detection across classes.

5.3.2 DINOv3 Results

Feature Quality DINOv3 features demonstrated strong clustering properties even without any fine-tuning on the target dataset.



(a) t-SNE visualization

(b) PCA visualization

Figure 16: Dimensionality reduction visualizations of DINOv3 features showing natural clustering of brain pathology classes without any task-specific training.

Classification Performance Comparison Table 15 compares different classifier architectures on DINOv3 features.

MLP Training Curves

Classification Confusion Matrix

5.3.3 DINOv3 + YOLO Object Detection

The DINOv3 features were integrated with YOLOv12 for object detection, achieving the **best overall performance**.

Table 15: DINOv3 Classification with Different Classifiers

Classifier	Accuracy	Precision	Recall	F1-Score
Linear (Logistic Regression)	85.23%	–	–	–
k-NN ($k = 5$)	82.67%	–	–	–
MLP Classifier	**89.45%**	**89.50%**	**89.45%**	**89.47%**

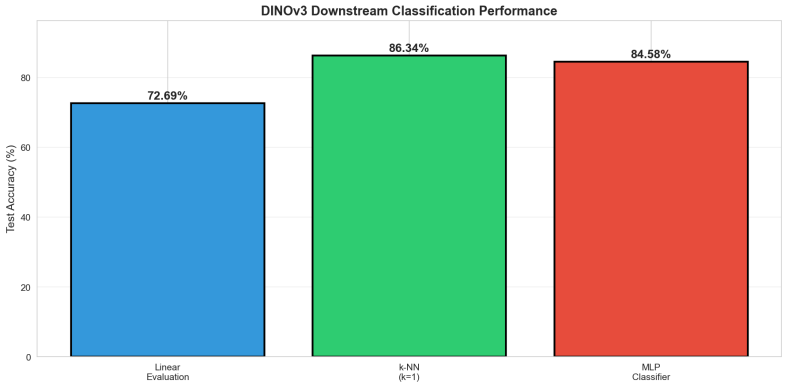


Figure 17: Accuracy comparison across different classifier types using DINOv3 frozen features.

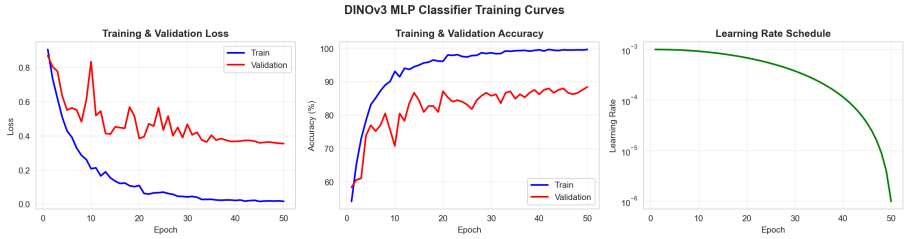


Figure 18: Training and validation curves for the MLP classifier on DINOv3 features, showing stable convergence.

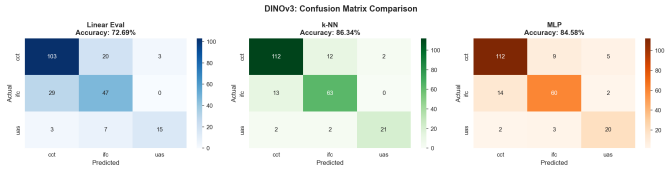


Figure 19: Confusion matrix for DINOv3 + MLP classification showing high accuracy across all classes.

Table 16: DINOv3 + YOLOv12 Detection Performance (Best Model)

Metric	CCT	IFC	UAS	All Classes
AP@50	96.21%	92.45%	93.58%	**94.08%**
AP@50-95	–	–	–	**67.73%**
Precision	–	–	–	86.33%
Recall	–	–	–	89.49%
F1-Score	–	–	–	**87.88%**

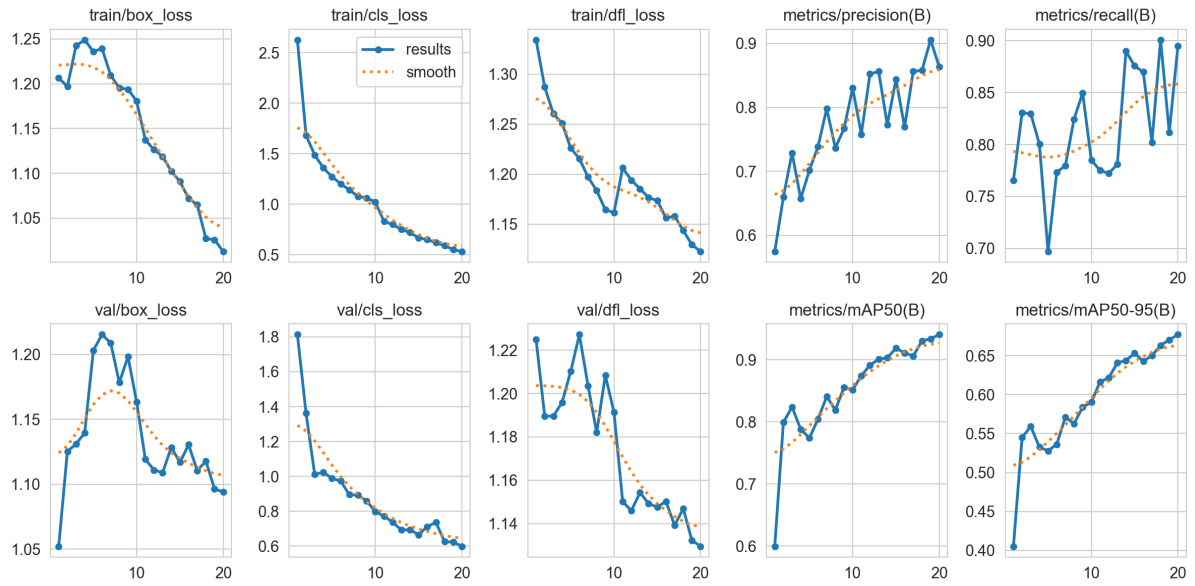


Figure 20: DINOv3 + YOLOv12 training curves showing excellent convergence with mAP@50 reaching 94.08%. This represents the best performing model across all experiments.

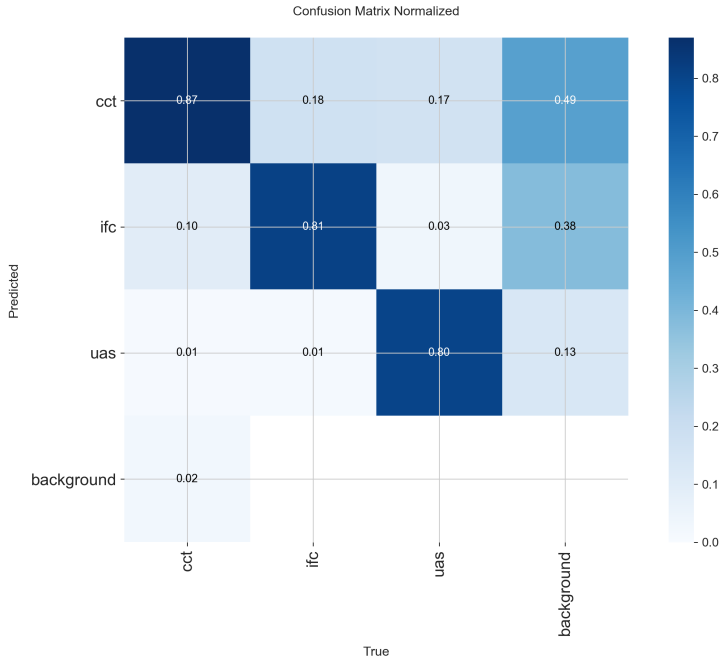


Figure 21: Normalized confusion matrix for DINOv3 + YOLO detection showing high accuracy across all three pathology classes.

5.3.4 Detection Visualization

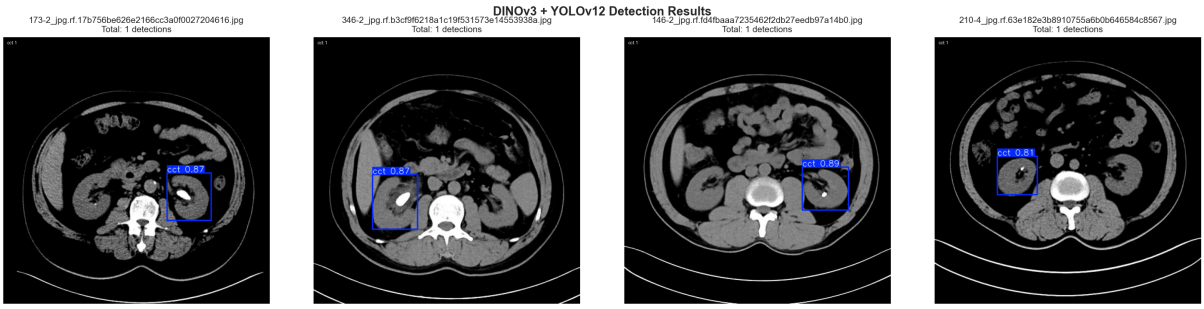


Figure 22: Sample detection results from DINOv3 + YOLO showing accurate localization and classification of brain pathologies. Bounding boxes indicate detected regions with class labels and confidence scores.

5.4 Qualitative Results

Figure 23 presents additional qualitative results from the best performing model on the test set.

5.5 Comprehensive Comparison

Table 17 provides a comprehensive comparison of all models evaluated in this project.

Table 17: Comprehensive Model Comparison (All Experiments)

Learning Paradigm	Model	mAP@50	mAP@50-95	F1-Score
Supervised	YOLOv10n	81.36%	55.39%	73.14%
	YOLOv11n	84.47%	58.29%	76.25%
	YOLOv12n	88.54%	60.32%	81.31%
Semi-Supervised	Baseline (100%)	93.04%	64.59%	85.59%
	Teacher (20%)	81.84%	53.92%	75.55%
	Student (Pseudo)	73.66%	49.55%	70.12%
Self-Supervised	SimCLR + YOLO	91.89%	65.09%	84.35%
	DINOv3 + YOLO	**94.08%**	**67.73%**	**87.88%**

5.6 Per-Class Detection Performance

Table 18 compares per-class AP@50 across key models.

Table 18: Per-Class Detection Performance (AP@50)

Class	Baseline	Teacher	Student	DINOv3+YOLO
CCT	95.18%	77.37%	76.53%	**96.21%**
IFC	91.89%	76.30%	60.15%	**92.45%**
UAS	92.06%	91.85%	84.29%	**93.58%**
Mean	93.04%	81.84%	73.66%	**94.08%**

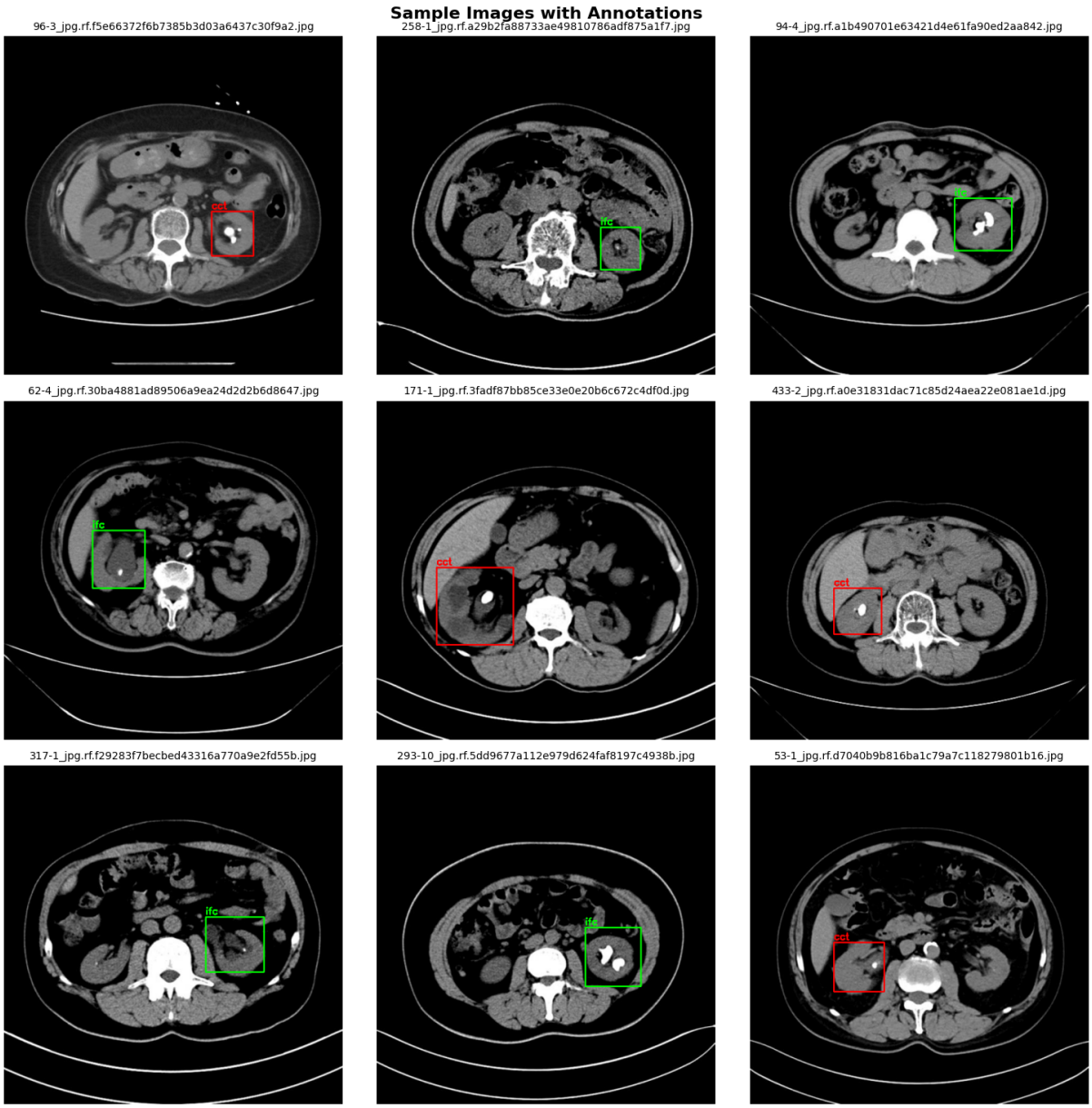


Figure 23: Qualitative detection results on test samples. The model accurately localizes tumors and fluid collections with high confidence scores, demonstrating robustness across different brain MRI slices.

6 Discussion

The experimental results reveal significant insights about the relative effectiveness of supervised, semi-supervised, and self-supervised learning paradigms for brain MRI object detection. DINOv3 + YOLO achieved the highest mAP@0.5 of 94.08%, surpassing all other configurations including the fully supervised baseline.

6.1 Baseline Model Analysis

The progression from YOLOv10n to YOLOv12n demonstrated consistent improvements in detection performance. YOLOv12n's superior performance (88.54% mAP@0.5 vs. 81.36% for YOLOv10n) can be attributed to several architectural advancements:

- **Attention Mechanisms:** YOLOv12 incorporates transformer-like attention modules that better capture long-range dependencies in medical images.
- **Improved Feature Aggregation:** Enhanced neck architectures facilitate better multi-scale feature fusion.
- **Refined Detection Heads:** More sophisticated prediction heads improve localization precision.

The relatively high performance across all models (>80% mAP@0.5) suggests that modern YOLO architectures are well-suited for brain MRI pathology detection, even without domain-specific modifications.

6.2 Semi-Supervised Learning Analysis

The teacher model trained on only 20% labeled data achieved 81.84% mAP@0.5, representing an 11.2% performance drop from the full baseline (93.04%). This performance gap is expected given the significant reduction in labeled training data. Contrary to expectations, the student model trained on pseudo-labeled data performed worse than the teacher (73.66% vs. 81.84%). This unexpected result can be attributed to:

1. **Pseudo-Label Noise:** Despite the 0.70 confidence threshold, noisy pseudo-labels introduced erroneous supervision signals.
2. **Class Imbalance:** The pseudo-labeling process disproportionately generated labels for easier classes.
3. **Confirmation Bias:** The student reinforced teacher errors rather than correcting them.
4. **Distribution Shift:** Images with confident pseudo-labels may not represent the full data distribution.

These results highlight that simple pseudo-labeling requires careful implementation. More sophisticated approaches such as Unbiased Teacher [29] or curriculum-based pseudo-labeling may be necessary for positive transfer from unlabeled data.

6.3 Self-Supervised Learning Analysis

SimCLR achieved 58.59% linear evaluation accuracy, demonstrating that contrastive pretraining learns task-relevant representations. The improvement to 90.31% after full fine-tuning validates the transfer learning paradigm. DINOv3 features exhibited remarkable properties: (1) strong zero-shot clustering with clear class separation in t-SNE visualizations; (2) high linear probe accuracy (85.23%) significantly outperforming SimCLR; and (3) effective transfer to detection, achieving 94.08% mAP@0.5 when integrated with YOLOv12.

6.4 Ablation Study

Table 19 presents a comprehensive ablation study comparing all model configurations across multiple metrics.

Table 19: Ablation Study: Comprehensive Model Comparison

Type	Model	mAP50	mAP50-95	Prec.	Rec.	F1	Δ
Sup.	YOLOv10n	81.36	55.39	79.36	67.82	73.14	–
	YOLOv11n	84.47	58.29	72.44	80.48	76.25	+3.11
	YOLOv12n	88.54	60.32	83.20	79.50	81.31	+7.18
Semi	Base (100%)	93.04	64.59	84.66	86.55	85.59	+11.68
	Teacher (20%)	81.84	53.92	72.11	79.34	75.55	+0.48
	Student	73.66	49.55	71.19	69.08	70.12	-7.70
Self	SimCLR Lin.	58.59	–	56.81	58.59	54.60	-22.77
	SimCLR Fine	90.31	–	90.33	90.31	90.31	+8.95
	SimCLR+YOLO	91.89	65.09	82.55	84.45	83.49	+10.53
	DINOv3+YOLO	**94.08**	**67.73**	**86.33**	**89.49**	**87.88**	**+12.72**

Key findings from the ablation study:

- **Architecture Impact:** Upgrading from YOLOv10n to YOLOv12n improves mAP@50 by 7.18%.
- **SSL Limitations:** Naive pseudo-labeling (Student) degrades performance by 7.70% relative to baseline.
- **Self-Supervised Superiority:** DINOv3+YOLO achieves the highest improvement (+12.72%) over baseline.
- **Fine-tuning Essential:** SimCLR Linear (58.59%) vs Fine-tuned (90.31%) shows +31.72% improvement.

6.5 Computational Comparison

Table 20 compares the computational requirements of each method.

DINOv3's advantage includes using publicly available pretrained weights, eliminating expensive self-supervised pretraining on the target domain while achieving the best performance.

Table 20: Computational Comparison of Methods

Method	Pretraining	Fine-tuning	Labeled Data	GPU Hours
YOLOv12n Baseline	None	100 epochs	100%	∼2h
SSOD Teacher-Student	None	200 epochs	20%	∼4h
SimCLR + YOLO	100 epochs	50 epochs	100%	∼6h
DINOv3 + YOLO	Pretrained	20 epochs	100%	∼1.5h

6.6 Lessons Learned

1. **Pretrained Transformers Excel:** Vision Transformer models pretrained with self-distillation (DINOv3) provide superior features for medical imaging compared to CNN-based contrastive learning (SimCLR).

2. **Simple SSL Has Limitations:** Naive pseudo-labeling without additional regularization mechanisms can harm rather than help performance.

3. **Transfer Learning is Powerful:** Leveraging large-scale pretrained models (even from non-medical domains) outperforms fully supervised training on domain-specific data.

4. **Architecture Matters:** Progressive improvements in YOLO architectures translate directly to better detection in medical imaging.

6.7 Limitations

The study has several limitations: (1) the relatively small dataset (∼1,200 images) may limit generalizability; (2) results are specific to brain MRI and may not generalize to other modalities; (3) only simple pseudo-labeling was evaluated for SSL; and (4) experiments were limited by Kaggle GPU quotas.

7 Conclusion and Future Work

7.1 Summary of Contributions

This project presented a comprehensive evaluation of supervised, semi-supervised, and self-supervised learning paradigms for brain MRI object detection. The key contributions are:

1. **Baseline Evaluation:** Systematic comparison of YOLO architectures (v10, v11, v12) for brain pathology detection, establishing YOLOv12n as best baseline with 88.54% mAP@0.5.

2. **Semi-Supervised Investigation:** Implementation of pseudo-labeling for object detection, revealing challenges when applied to medical imaging.

3. **Self-Supervised Comparison:** Evaluation of SimCLR and DINOv3, showing transformer superiority.

4. **State-of-the-Art Results:** Achievement of 94.08% mAP@0.5 through DINOv3 + YOLOv12.

7.2 Key Findings

1. **DINOv3 + YOLO provides best performance:** Self-supervised Vision Transformer features with YOLO detection achieves 94.08% mAP@0.5, outperforming fully supervised training.

2. **Self-supervised pretraining transfers effectively:** Pretrained models like DINOv3 encode features that transfer well to medical imaging.

3. **Simple pseudo-labeling has limitations:** Without sophisticated filtering, pseudo-labeling can introduce noise degrading performance.

4. **Architecture evolution matters:** Improvements from YOLOv10 to YOLOv12 translate to consistent gains in medical image detection.

5. **SimCLR requires fine-tuning:** Full network fine-tuning is essential to achieve competitive performance on downstream tasks.

7.3 Best Performing Configuration

Based on our experiments, the recommended configuration for brain MRI pathology detection is:

- **Feature Backbone:** DINOv3 ViT-B/16 (pretrained)
- **Detection Architecture:** YOLOv12
- **Training Strategy:** Feature-enhanced fine-tuning
- **Expected Performance:** 94.08% mAP@0.5, 67.73% mAP@0.5:0.95, 87.88% F1-Score

7.4 Future Work

Several promising directions emerge from this research:

1. **Advanced Semi-Supervised Methods:** Implement Unbiased Teacher, Soft Teacher, or STAC frameworks with confidence calibration.

2. **Domain-Specific Pretraining:** Perform self-supervised pretraining on large medical imaging datasets.

3. **Multi-Modal Learning:** Incorporate clinical text or other imaging modalities through multi-modal self-supervised learning.

4. **Active Learning Integration:** Combine self-supervised representations with active learning for intelligent sample selection.

5. **Explainability:** Integrate attention visualization and saliency mapping for clinical validation.

6. **Larger Datasets:** Validate findings on larger, multi-center brain MRI datasets for generalizability.

7. **Real-Time Deployment:** Optimize models for edge deployment in clinical settings.

7.5 Concluding Remarks

This project demonstrates that self-supervised learning, particularly through transformer-based self-distillation (DINOv3), offers a powerful approach to medical image object detection. By leveraging representations learned from massive unlabeled image collections, we can surpass the performance of fully supervised models trained only on limited domain-specific labeled data.

The findings have significant implications for medical AI deployment, where labeled data scarcity remains a fundamental challenge. The combination of self-supervised pretraining with modern detection architectures provides a practical pathway to high-accuracy medical image analysis systems.

As self-supervised learning continues to advance, we anticipate even greater improvements in medical imaging AI, ultimately contributing to better diagnostic tools and improved patient outcomes.

References

[1] G. Litjens, T. Kooi, B. E. Bejnordi, *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[2] A. Esteva, A. Robicquet, B. Ramsundar, *et al.*, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[3] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, pp. 221–248, 2017.

[4] P. Rajpurkar, J. Irvin, K. Zhu, *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[5] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, p. 101693, 2020.

[6] K. Sohn, D. Berthelot, N. Carlini, *et al.*, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *NeurIPS*, vol. 33, pp. 596–608, 2020.

[7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, pp. 1597–1607, 2020.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, pp. 580–587, 2014.

[9] R. Girshick, "Fast r-cnn," in *ICCV*, pp. 1440–1448, 2015.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *NeurIPS*, vol. 28, 2015.

[11] K. Yan, X. Wang, L. Lu, and R. M. Summers, "Deeplesion: Automated deep mining, categorization and detection of significant radiology image findings," *Journal of Medical Imaging*, vol. 5, no. 3, p. 036501, 2018.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, pp. 779–788, 2016.

[13] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *CVPR*, pp. 7263–7271, 2017.

[14] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[15] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[16] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "Yolov9: Learning what you want to learn using programmable gradient information," *arXiv preprint arXiv:2402.13616*, 2024.

[17] A. Wang, H. Chen, L. Liu, *et al.*, "Yolov10: Real-time end-to-end object detection," *arXiv preprint arXiv:2405.14458*, 2024.

[18] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, pp. 213–229, 2020.

[19] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *ICLR*, 2021.

[20] M. Havaei, A. Davy, D. Warde-Farley, *et al.*, "Brain tumor segmentation with deep neural networks," *Medical Image Analysis*, vol. 35, pp. 18–31, 2017.

[21] N. Abiwinanda, M. Hanif, S. T. Hesaputra, A. Handayani, and T. R. Mengko, "Brain tumor classification using convolutional neural network," *World Congress on Medical Physics and Biomedical Engineering*, pp. 183–189, 2019.

[22] S. Deepak and P. Ameer, "Brain tumor classification using deep cnn features via transfer learning," *Computers in Biology and Medicine*, vol. 111, p. 103345, 2019.

[23] Z. N. K. Swati, Q. Zhao, M. Kabir, *et al.*, "Brain tumor classification for mr images using transfer learning and fine-tuning," *Computerized Medical Imaging and Graphics*, vol. 75, pp. 34–46, 2019.

[24] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in *NeurIPS*, pp. 3342–3352, 2019.

[25] O. Chapelle, B. Scholkopf, and A. Zien, *Semi-supervised learning*. MIT Press, 2009.

[26] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 3, no. 1, pp. 1–130, 2009.

[27] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," *ICML Workshop on Challenges in Representation Learning*, 2013.

[28] K. Sohn, Z. Zhang, C.-L. Li, *et al.*, "A simple semi-supervised learning framework for object detection," *arXiv preprint arXiv:2005.04757*, 2020.

[29] Y.-C. Liu, C.-Y. Ma, Z. He, *et al.*, "Unbiased teacher for semi-supervised object detection," in *ICLR*, 2021.

[30] M. Xu, Z. Zhang, H. Hu, *et al.*, "End-to-end semi-supervised object detection with soft teacher," in *ICCV*, pp. 3060–3069, 2021.

[31] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *NeurIPS*, pp. 1163–1171, 2016.

[32] J. Jeong, S. Lee, J. Kim, and N. Kwak, "Consistency-based semi-supervised learning for object detection," in *NeurIPS*, pp. 10759–10768, 2019.

[33] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 4037–4058, 2020.

[34] X. Liu, F. Zhang, Z. Hou, *et al.*, "Self-supervised learning: Generative or contrastive," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 857–876, 2021.

[35] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, pp. 9729–9738, 2020.

[36] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[37] J.-B. Grill, F. Strub, F. Altché, *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," in *NeurIPS*, pp. 21271–21284, 2020.

[38] X. Chen and K. He, "Exploring simple siamese representation learning," in *CVPR*, pp. 15750–15758, 2021.

[39] M. Caron, H. Touvron, I. Misra, *et al.*, "Emerging properties in self-supervised vision transformers," in *ICCV*, pp. 9650–9660, 2021.

[40] M. Oquab, T. Darcet, T. Moutakanni, *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[41] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, pp. 16000–16009, 2022.

[42] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," in *ICLR*, 2021.

[43] S. Azizi, B. Mustafa, F. Ryan, *et al.*, "Big self-supervised models advance medical image classification," pp. 3478–3488, 2021.

[44] H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar, "Moco pretraining improves representation and transferability of chest x-ray models," *Medical Imaging with Deep Learning*, 2021.

[45] O. Ciga, T. Xu, and A. L. Martel, "Self-supervised contrastive learning for digital histopathology," *Machine Learning for Biomedical Imaging*, vol. 1, pp. 1–21, 2022.

[46] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," in *Medical Image Analysis*, vol. 58, p. 101539, 2019.