

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	7
1.1	Background and Motivation	7
1.2	The Challenge of Limited Labeled Data	7
1.3	Project Objectives	7
1.4	Dataset Overview	7
1.5	Report Organization	8
2	Literature Review	8
2.1	Object Detection for Medical Imaging	8
2.1.1	Region-Based Methods	8
2.1.2	Single-Stage Detectors	8
2.1.3	Transformer-Based Detection	9
2.2	Brain MRI Analysis and Tumor Detection	9
2.3	Semi-Supervised Learning for Object Detection	9
2.3.1	Pseudo-Labeling	9
2.3.2	Teacher-Student Frameworks	9
2.3.3	Consistency Regularization	9
2.4	Self-Supervised Representation Learning	9
2.4.1	Contrastive Learning	10
2.4.2	Self-Distillation Methods	10
2.4.3	Vision Transformers for Self-Supervised Learning	10
2.4.4	Masked Image Modeling	10
2.5	Self-Supervised Learning in Medical Imaging	10
2.6	Gap Analysis and Contributions	10
3	Methodology	11
3.1	Dataset	11
3.1.1	Dataset Description	11
3.1.2	Class Definitions	11
3.1.3	Class Distribution	11
3.1.4	Data Splitting	12
3.2	Supervised Learning Baseline	12
3.2.1	YOLO Architecture Family	12
3.3	Semi-Supervised Object Detection	13
3.3.1	Pseudo-Labeling Framework	13
3.3.2	Key Hyperparameters	14
3.4	Self-Supervised Learning Methods	14
3.4.1	SimCLR: Contrastive Learning	14
3.4.2	DINOv3: Self-Distillation with Vision Transformers	15
3.5	YOLO Integration for Object Detection	16
3.6	Evaluation Metrics	17

4 Experimental Setup	17
4.1 Computational Environment	17
4.2 Experiment Groups	17
4.2.1 Experiment Group 1: Supervised Learning Baseline	18
4.2.2 Experiment Group 2: Semi-Supervised Object Detection	18
4.2.3 Experiment Group 3: Self-Supervised Learning	19
4.3 Training Procedures	20
4.3.1 Loss Functions	20
4.3.2 Early Stopping and Checkpointing	20
4.4 Evaluation Protocol	21
5 Results	21
5.1 Supervised Learning Baseline Results	21
5.1.1 Model Comparison	21
5.1.2 Training Curves	21
5.1.3 Confusion Matrix Analysis	23
5.1.4 Precision-Recall Curves	23
5.2 Semi-Supervised Learning Results	23
5.2.1 Teacher-Student Performance	23
5.2.2 Model Comparison Visualization	24
5.2.3 Pseudo-Label Quality Analysis	24
5.2.4 Baseline Training Curves	24
5.3 Self-Supervised Learning Results	24
5.3.1 SimCLR Results	24
5.3.2 DINOv3 Results	28
5.3.3 DINOv3 + YOLO Object Detection	29
5.3.4 Detection Visualization	29
5.4 Qualitative Results	29
5.5 Comprehensive Comparison	33
5.6 Per-Class Detection Performance	33
6 Discussion	33
6.1 Baseline Model Analysis	33
6.2 Semi-Supervised Learning Analysis	34
6.2.1 Teacher Model Performance	34
6.2.2 Student Model Underperformance	34
6.2.3 Implications for SSL in Medical Imaging	34
6.3 Self-Supervised Learning Analysis	34
6.3.1 SimCLR: Contrastive Learning	34
6.3.2 DINOv3: Self-Distillation with Vision Transformers	35
6.4 Comparative Analysis	35
6.4.1 Best Overall Performance	35
6.4.2 Label Efficiency	35

6.4.3	Computational Considerations	35
6.5	Lessons Learned	36
6.6	Limitations	36
6.7	Practical Implications	36
7	Conclusion and Future Work	37
7.1	Summary of Contributions	37
7.2	Key Findings	37
7.3	Best Performing Configuration	37
7.4	Future Work	38
7.5	Concluding Remarks	38

List of Figures

1	Baseline model comparison showing performance metrics for YOLOv10n, YOLOv11n, and YOLOv12n on the Brain MRI dataset.	13
2	SimCLR augmentation pairs showing different views of the same brain MRI image used for contrastive learning.	15
3	t-SNE visualization of DINOv3 features showing natural clustering of brain MRI classes in the learned feature space.	16
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.	22
5	Normalized confusion matrix for YOLOv12n baseline model showing per-class detection accuracy. High diagonal values indicate strong class discrimination.	22
6	Precision-Recall curves for YOLOv12n across all three classes, demonstrating high area under the curve for each pathology type.	23
7	Comparison of Teacher vs Student model performance in the semi-supervised learning framework.	24
8	Analysis of pseudo-label quality showing confidence distribution and detection coverage on unlabeled data.	24
9	Training curves for the SSOD baseline model (100% labeled data) showing loss components and mAP evolution.	25
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.	25
11	t-SNE visualization of SimCLR features after full fine-tuning, showing clear separation between the three brain pathology classes.	26
12	Confusion matrices comparing Linear Evaluation (left) versus Full Fine-tuning (right). Full fine-tuning achieves near-diagonal matrices indicating high classification accuracy.	27
13	Per-class precision, recall, and F1-score for SimCLR classification, showing balanced performance across all three pathology classes.	27
14	SimCLR + YOLOv12 detection results	27
15	SimCLR + YOLO training curves and confusion matrix. The model demonstrates good convergence and balanced detection across classes.	28
16	Dimensionality reduction visualizations of DINOv3 features showing natural clustering of brain pathology classes without any task-specific training.	28
17	Accuracy comparison across different classifier types using DINOv3 frozen features.	29
18	Training and validation curves for the MLP classifier on DINOv3 features, showing stable convergence.	29
19	Confusion matrix for DINOv3 + MLP classification showing high accuracy across all classes.	29
20	DINOv3 + YOLOv12 training curves showing excellent convergence with mAP@50 reaching 94.08%. This represents the best performing model across all experiments.	30

21	Normalized confusion matrix for DINOv3 + YOLO detection showing high accuracy across all three pathology classes.	31
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.	31
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.	32

List of Tables

1	Dataset Overview	11
2	Class Distribution	12
3	Data Split Configuration	12
4	Semi-Supervised Learning Configuration	14
5	Computational Environment	17
6	Baseline Experiment Configuration	18
7	Semi-Supervised Experiment Configuration	18
8	SimCLR Pretraining Configuration	19
9	SimCLR Fine-tuning Configuration	19
10	DINOv3 Feature Extraction and Classification	20
11	Supervised Learning Baseline Performance Comparison	21
12	Semi-Supervised Object Detection Results	23
13	SimCLR Classification Performance	24
14	DINOv3 Classification with Different Classifiers	28
15	DINOv3 + YOLOv12 Detection Performance (Best Model)	30
16	Comprehensive Model Comparison (All Experiments)	33
17	Per-Class Detection Performance (AP@50)	33
18	Computational Comparison of Methods	35

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

This section reviews the relevant literature on object detection, medical image analysis, semi-supervised learning, and self-supervised learning methods that form the theoretical foundation of this project.

2.1 Object Detection for Medical Imaging

Object detection in medical imaging has evolved rapidly with the advent of deep learning. Early approaches relied on handcrafted features combined with traditional machine learning classifiers, but these have been largely superseded by end-to-end deep learning methods [1].

2.1.1 Region-Based Methods

Region-based Convolutional Neural Networks (R-CNN) and their variants introduced a two-stage approach: generating region proposals 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 and tumor localization [11].

2.1.2 Single-Stage Detectors

Single-stage detectors eliminate the region proposal stage, directly predicting bounding boxes and class probabilities. YOLO (You Only Look Once) pioneered this approach, framing detection as a regression problem [12]. Subsequent versions improved accuracy and speed: YOLOv2 introduced batch normalization and anchor boxes [13], YOLOv3 added multi-scale predictions [14], and YOLOv4-v9 incorporated various architectural improvements [15, 16].

Recent YOLO variants (YOLOv10, YOLOv11, YOLOv12) have introduced attention mechanisms and improved feature pyramid networks, achieving state-of-the-art performance on standard benchmarks while maintaining real-time inference capabilities [17].

2.1.3 Transformer-Based Detection

The DETR (DEtection TRansformer) architecture introduced transformers to object detection, eliminating the need for hand-designed components like anchor boxes [18]. Subsequent works improved training efficiency and small object detection performance [19].

2.2 Brain MRI Analysis and Tumor Detection

Automated analysis of brain MRI has received significant attention due to its clinical importance. Deep learning approaches have demonstrated promising results for brain tumor segmentation [20], classification [21], and detection [22].

Convolutional Neural Networks have been applied to classify brain tumors into categories such as glioma, meningioma, and pituitary tumors [23]. Transfer learning from ImageNet-pretrained models has proven particularly effective for medical imaging tasks with limited data [24].

2.3 Semi-Supervised Learning for Object Detection

Semi-supervised learning addresses the challenge of limited labeled data by leveraging unlabeled samples during training [25, 26].

2.3.1 Pseudo-Labeling

Pseudo-labeling generates labels for unlabeled data using model predictions, then retrains the model on the expanded dataset [27]. For object detection, this involves generating bounding box predictions with confidence thresholds to filter low-quality pseudo-labels [28].

2.3.2 Teacher-Student Frameworks

Teacher-student frameworks use a teacher model to generate pseudo-labels for training a student model. The Unbiased Teacher method addresses pseudo-label bias in object detection [29]. STAC (Self-Training with Augmented Consistency) combines pseudo-labeling with strong data augmentation [28]. Soft Teacher relaxes hard pseudo-label assignments through soft weighting schemes [30].

2.3.3 Consistency Regularization

Consistency regularization enforces prediction consistency under different perturbations of the same input [31]. FixMatch combines pseudo-labeling with consistency regularization using weak and strong augmentations [6]. These principles have been extended to object detection tasks [32].

2.4 Self-Supervised Representation Learning

Self-supervised learning learns representations from unlabeled data through pretext tasks, enabling effective transfer to downstream tasks [33, 34].

2.4.1 Contrastive Learning

Contrastive learning learns representations by maximizing agreement between differently augmented views of the same image while minimizing agreement with other images [7]. SimCLR introduced a simple yet effective framework with strong data augmentation and learnable non-linear projections [7]. MoCo (Momentum Contrast) used a momentum-updated encoder and memory bank for efficient contrastive learning [35, 36].

2.4.2 Self-Distillation Methods

BYOL (Bootstrap Your Own Latent) demonstrated that contrastive learning could work without negative samples through momentum-based self-distillation [37]. SimSiam simplified this further by removing the momentum encoder [38].

2.4.3 Vision Transformers for Self-Supervised Learning

DINO (self-DIstillation with NO labels) applied self-distillation to Vision Transformers, achieving impressive results without using any labels [39]. DINOv2 scaled this approach to larger datasets and models, learning powerful visual features that generalize across tasks [40]. These features have proven particularly effective for transfer learning to specialized domains including medical imaging [41].

2.4.4 Masked Image Modeling

Masked Autoencoders (MAE) learn representations by reconstructing masked image patches, inspired by masked language modeling in NLP [42]. BEiT introduced vision tokenization for masked prediction [43].

2.5 Self-Supervised Learning in Medical Imaging

Self-supervised learning has shown particular promise in medical imaging where labeled data is scarce. Contrastive learning methods have been adapted for chest X-rays [44], histopathology [45], and radiology [41]. These methods often outperform ImageNet pretraining when transferred to medical tasks [46].

2.6 Gap Analysis and Contributions

While previous works have explored semi-supervised and self-supervised learning separately, few studies have systematically compared these approaches for medical object detection. This project addresses this gap by:

1. Providing a comprehensive comparison of baseline supervised detectors (YOLOv10/11/12) on brain MRI data
2. Implementing and evaluating pseudo-labeling for semi-supervised detection
3. Comparing two distinct self-supervised paradigms (contrastive vs. self-distillation)
4. Demonstrating the effectiveness of DINOv3 features for medical object detection

3 Methodology

This section describes the dataset, model architectures, and training methodologies employed in this project.

3.1 Dataset

3.1.1 Dataset Description

The experiments utilize a Brain MRI object detection dataset specifically curated for pathology detection¹. The dataset contains approximately 1,200 high-resolution MRI scans with expert annotations.

Table 1: 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, each representing a distinct type of brain abnormality:

- **CCT (Cerebral Cortex Tumor):** Tumorous masses originating in or affecting the cerebral cortex, typically appearing as enhanced regions on contrast MRI.
- **IFC (Intracerebral Fluid Collection):** Abnormal accumulations of cerebrospinal fluid or other fluids within the brain parenchyma, often indicating pathological conditions.
- **UAS (Unidentified Anomaly Signature):** Other anomalies that require clinical attention but do not fall into the specific CCT or IFC categories.

3.1.3 Class Distribution

The dataset exhibits a relatively balanced class distribution, which is beneficial for training unbiased models:

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

Table 2: Class Distribution

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%	

3.1.4 Data Splitting

The dataset was split into training, validation, and test sets with the following ratios:

Table 3: Data Split Configuration

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:

- **Labeled set (20%):** ~192 images with annotations
- **Unlabeled set (80%):** ~768 images (labels withheld)

3.2 Supervised Learning Baseline

3.2.1 YOLO Architecture Family

We evaluated three state-of-the-art YOLO architectures for the baseline supervised experiments:

YOLOv10n YOLOv10 introduced consistent dual assignments for NMS-free training and an efficiency-accuracy driven model design. The nano variant (YOLOv10n) provides a lightweight architecture suitable for resource-constrained deployments while maintaining competitive accuracy.

YOLOv11n YOLOv11 improved upon YOLOv10 with enhanced feature aggregation and refined anchor-free detection heads. It introduced improved CSPNet backbone modifications and better gradient flow for 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 the efficiency hallmarks of the YOLO family.

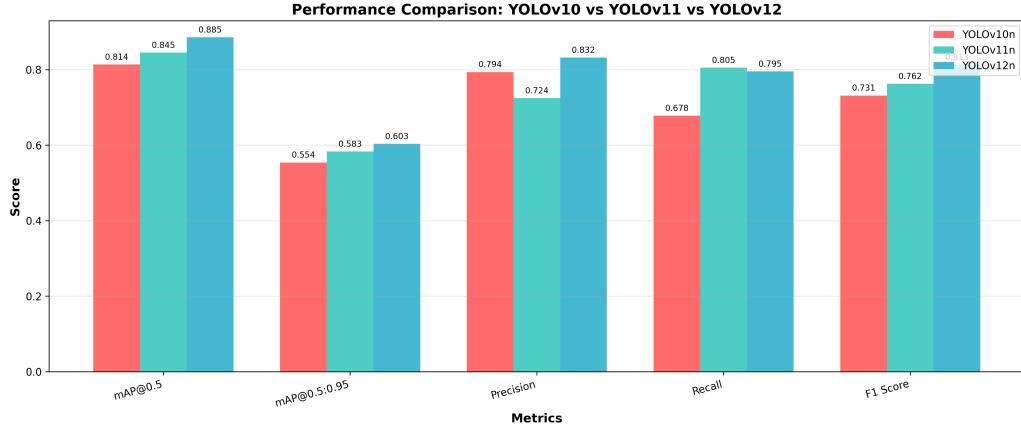


Figure 1: Baseline model comparison showing performance metrics for YOLOv10n, YOLOv11n, and YOLOv12n on the Brain MRI dataset.

3.3 Semi-Supervised Object Detection

3.3.1 Pseudo-Labeling Framework

The semi-supervised object detection pipeline employs a teacher-student framework with pseudo-labeling:

Algorithm 1 Pseudo-Labeling for Semi-Supervised Object Detection

Require: Labeled set D_L , Unlabeled set D_U , Confidence threshold τ

- 1: **Stage 1: Train Teacher Model**
 - 2: Train YOLOv12 on D_L for 100 epochs → Teacher weights W_T
 - 3: **Stage 2: Generate Pseudo-Labels**
 - 4: **for** each image $x \in D_U$ **do**
 - 5: $\hat{y} \leftarrow \text{Teacher}(x)$
 - 6: **if** confidence(\hat{y}) $\geq \tau$ **then**
 - 7: Add (x, \hat{y}) to pseudo-labeled set D_P
 - 8: **end if**
 - 9: **end for**
 - 10: **Stage 3: Train Student Model**
 - 11: Train YOLOv12 on $D_L \cup D_P$ for 100 epochs → Student weights W_S
 - 12: **return** W_S
-

3.3.2 Key Hyperparameters

Table 4: Semi-Supervised Learning Configuration

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 (Simple Contrastive Learning of Representations) learns visual representations by maximizing agreement between differently augmented views of the same image.

Architecture

- **Encoder:** ResNet-18 backbone
- **Projection Head:** 2-layer MLP (2048 → 256 → 128)
- **Output Dimension:** 128-dimensional embedding

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

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

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

Data Augmentation Pipeline

- Random resized crop (scale: 0.2–1.0)
- Random horizontal flip
- Color jittering (brightness, contrast, saturation, hue)
- Random grayscale (probability: 0.2)
- Gaussian blur

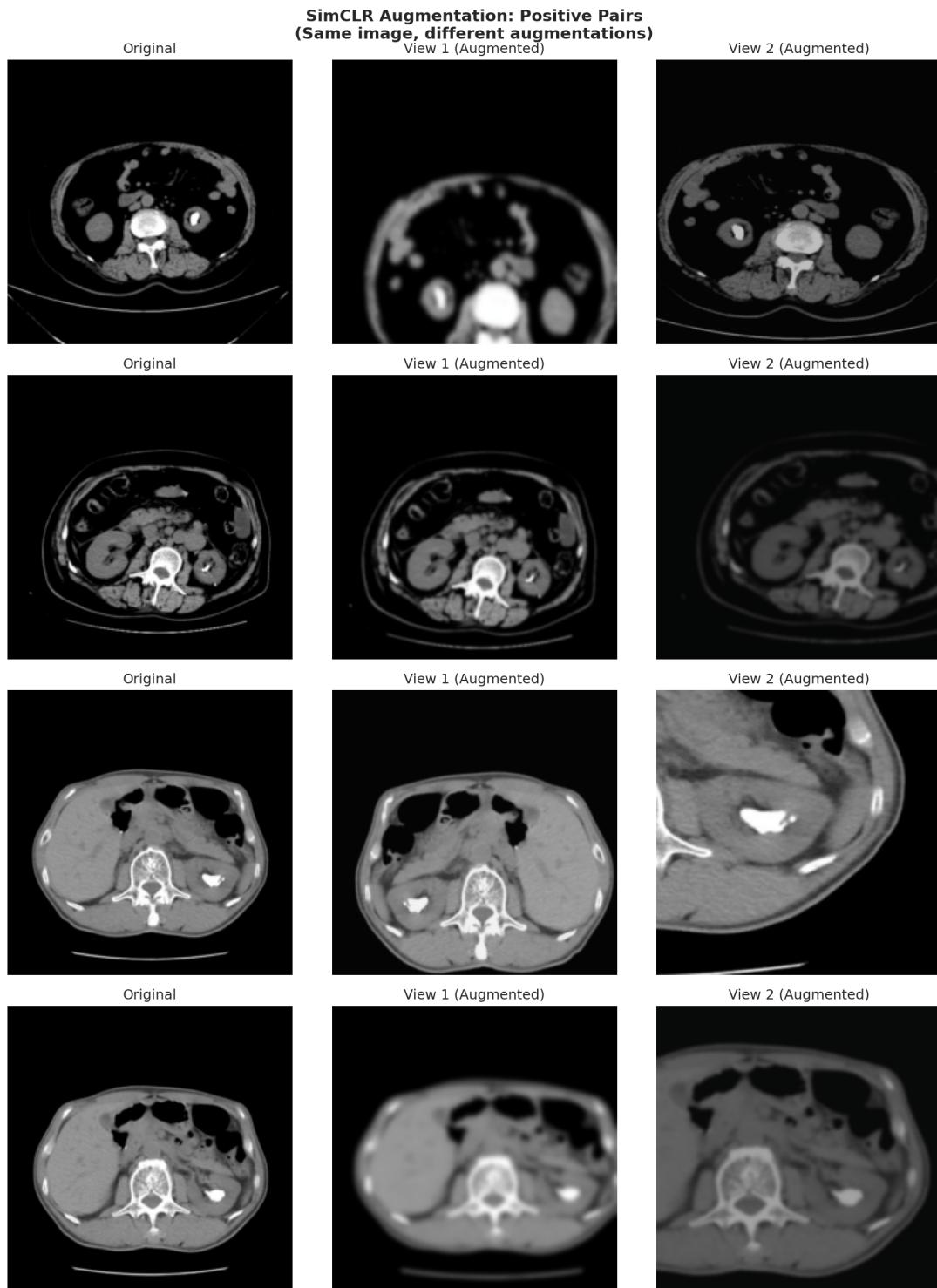


Figure 2: SimCLR augmentation pairs showing different views of the same brain MRI image used for contrastive learning.

3.4.2 DINOv3: Self-Distillation with Vision Transformers

DINOv3 employs self-distillation with Vision Transformers, learning powerful visual representations without any labels.

Architecture

- **Backbone:** Vision Transformer ViT-B/16
- **Parameters:** 86 million
- **Feature Dimension:** 768
- **Pretraining Data:** LVD-1689M (1.7 billion images)

Feature Extraction DINOv3 extracts features using the [CLS] token representation from the final transformer layer. For an input image x :

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

Downstream Classifiers Three classification approaches were evaluated:

- **Linear Classifier:** Logistic regression on frozen features
- **k-NN Classifier:** k -nearest neighbors ($k = 5$) in feature space
- **MLP Classifier:** $768 \rightarrow 256 \rightarrow 128 \rightarrow 3$ with ReLU activation

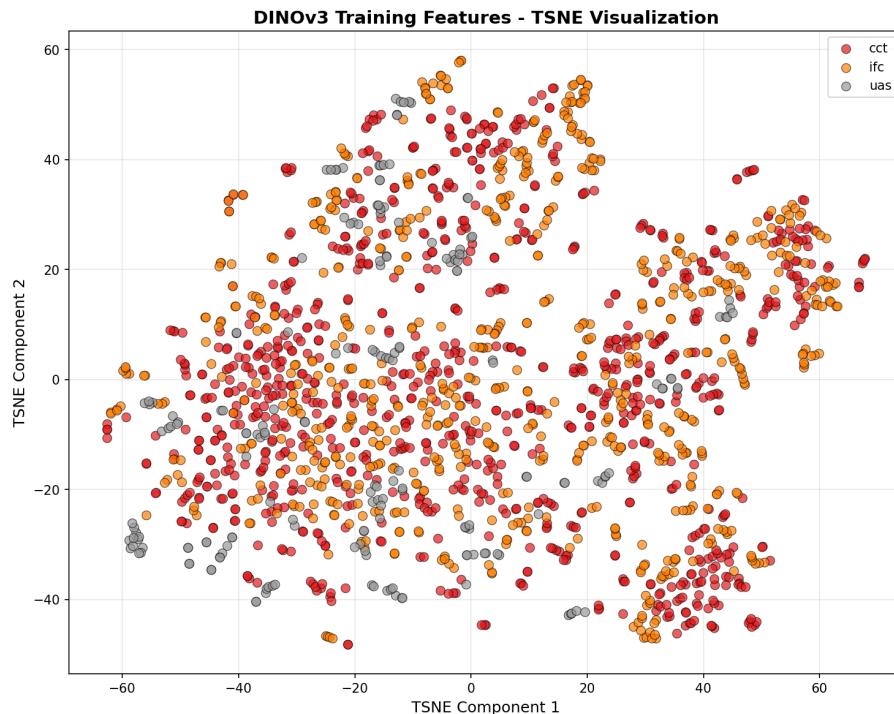


Figure 3: t-SNE visualization of DINOv3 features showing natural clustering of brain MRI classes in the learned feature space.

3.5 YOLO Integration for Object Detection

Both self-supervised methods were integrated with YOLOv12 for object detection:

1. **Feature-Enhanced Training:** Self-supervised features used to enhance training data or initialize components.

2. **Classification-Guided Detection:** Classification predictions used to improve detection confidence estimation.
3. **Fine-tuning Pipeline:** Transfer learned representations to the detection task through fine-tuning on labeled data.

3.6 Evaluation Metrics

The following metrics were used to evaluate model performance:

- **mAP@0.5:** Mean Average Precision at IoU threshold of 0.5
- **mAP@0.5:0.95:** Mean Average Precision averaged over IoU thresholds from 0.5 to 0.95
- **Precision:** Ratio of true positives to all positive predictions
- **Recall:** Ratio of true positives to all actual positives
- **F1-Score:** Harmonic mean of precision and recall

4 Experimental Setup

This section details the computational environment, training configurations, and experimental protocols used in this project.

4.1 Computational Environment

All experiments were conducted using cloud-based GPU computing resources:

Table 5: Computational Environment

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

The experiments were organized into three main groups corresponding to the project objectives:

4.2.1 Experiment Group 1: Supervised Learning Baseline

Table 6: Baseline Experiment Configuration

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

Data Augmentation (applied during training):

- Mosaic augmentation
- Random HSV shifts
- Random horizontal/vertical flips
- Random scaling and translation

4.2.2 Experiment Group 2: Semi-Supervised Object Detection

Table 7: Semi-Supervised Experiment Configuration

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

4.2.3 Experiment Group 3: Self-Supervised Learning

Table 8: 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

Table 9: SimCLR Fine-tuning Configuration

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

SimCLR Fine-tuning Configuration

Table 10: DINov3 Feature Extraction and Classification

Parameter	Value
<i>Feature Extraction</i>	
Pretrained Model	DINov3 ViT-B/16
Feature Dimension	768
Source	Facebook AI Research
<i>MLP Classifier</i>	
Architecture	$768 \rightarrow 256 \rightarrow 128 \rightarrow 3$
Activation	ReLU
Dropout	0.3
Epochs	100
Learning Rate	0.001
Optimizer	Adam
<i>YOLO Integration</i>	
Base Model	YOLOv12
Training Epochs	20
Batch Size	16

DINov3 Configuration

4.3 Training Procedures

4.3.1 Loss Functions

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

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

where \mathcal{L}_{box} is the CIoU loss for bounding box regression, \mathcal{L}_{cls} is the classification loss, and \mathcal{L}_{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:

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

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 11 summarizes the test set performance.

Table 11: 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

Key Observations:

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

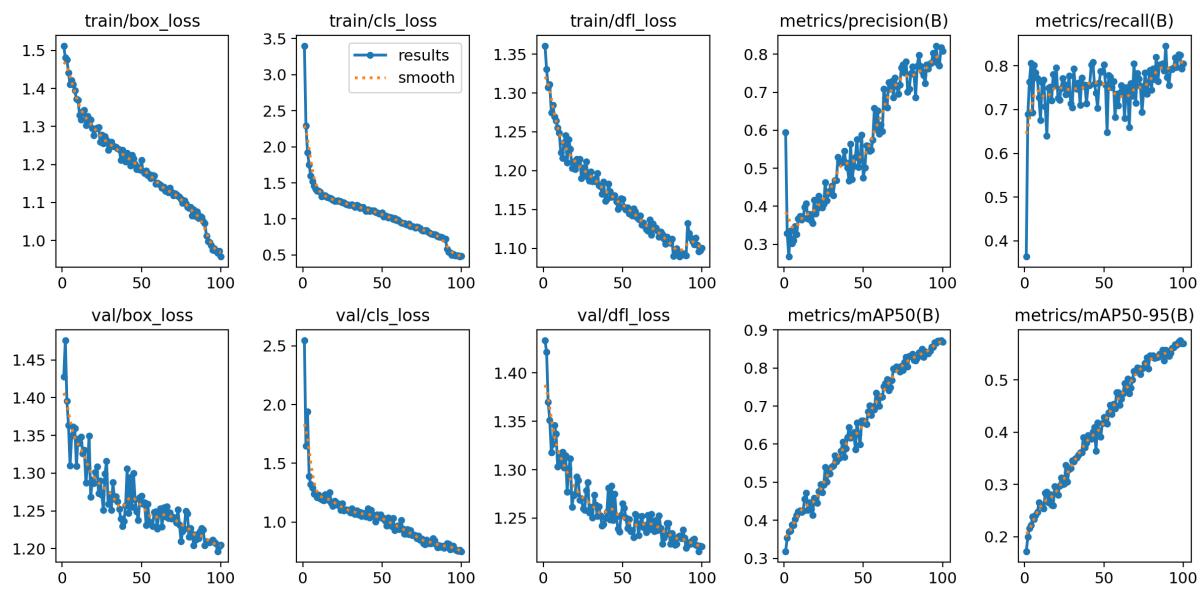


Figure 4: YOLOv12n training curves showing loss components (box, classification, objectness) and mAP progression over 100 epochs. The model shows stable convergence with consistent improvement in detection metrics.

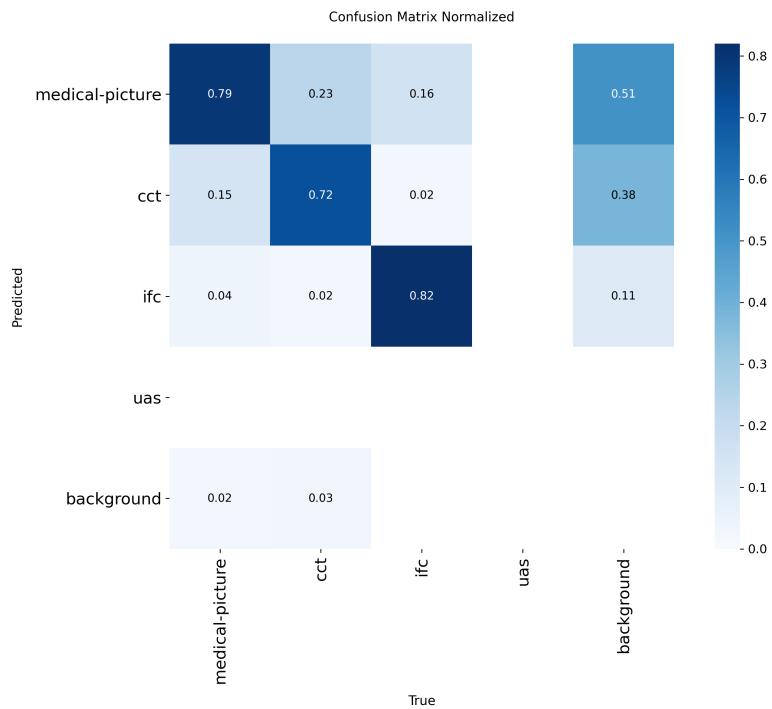


Figure 5: Normalized confusion matrix for YOLOv12n baseline model showing per-class detection accuracy. High diagonal values indicate strong class discrimination.

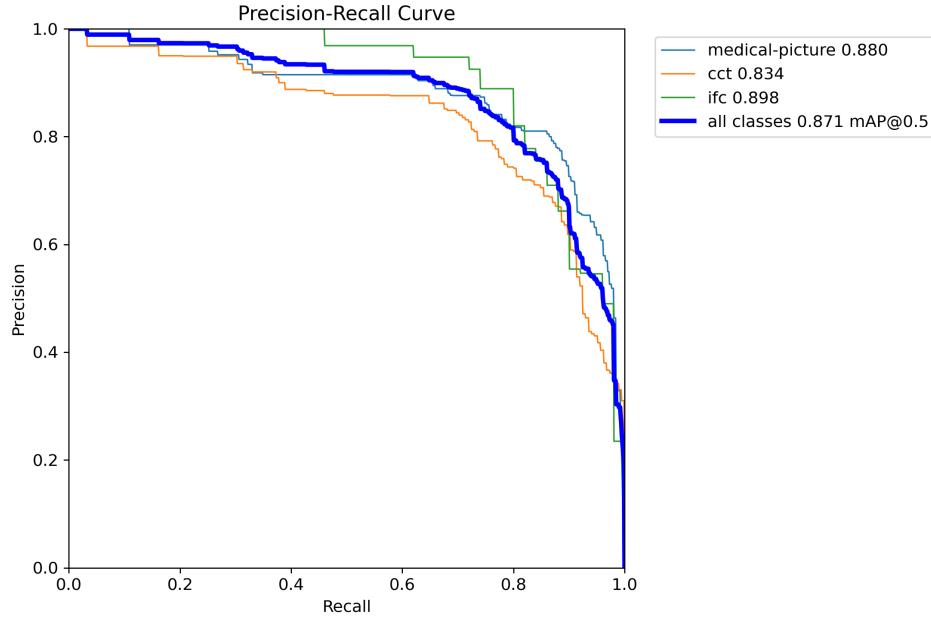


Figure 6: Precision-Recall curves for YOLOv12n across all three classes, demonstrating high area under the curve for each pathology type.

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 12 compares the semi-supervised detection results against the baseline.

Table 12: 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%

Key Observations:

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

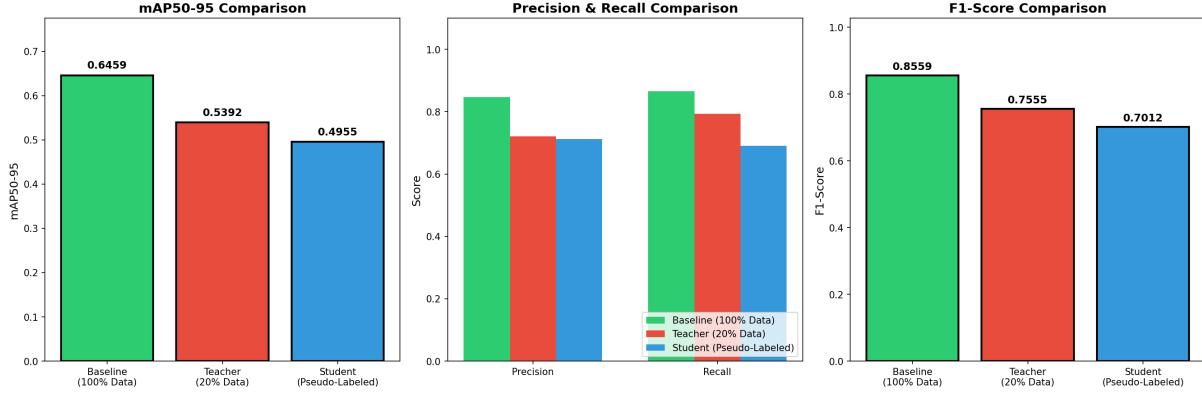


Figure 7: Comparison of Teacher vs Student model performance in the semi-supervised learning framework.

5.2.2 Model Comparison Visualization

5.2.3 Pseudo-Label Quality Analysis

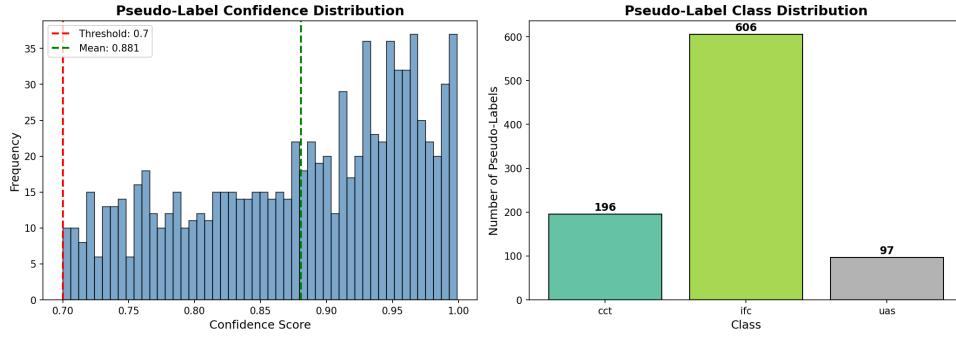


Figure 8: Analysis of pseudo-label quality showing confidence distribution and detection coverage on unlabeled data.

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.

Classification Performance Table 13 compares linear evaluation and full fine-tuning performance on the classification task.

Table 13: 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%

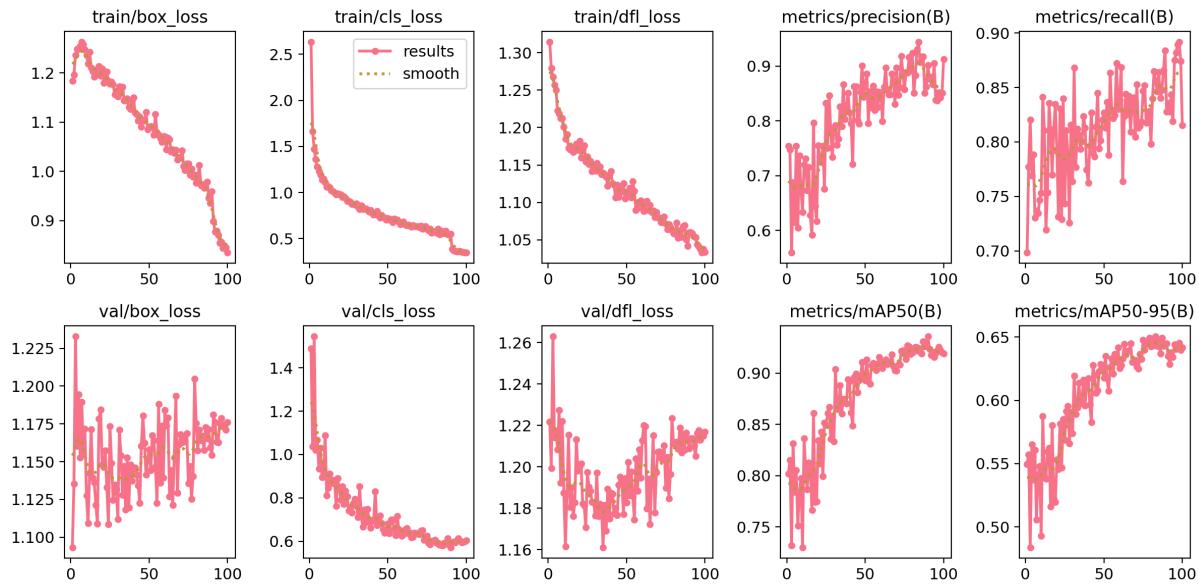


Figure 9: Training curves for the SSOD baseline model (100% labeled data) showing loss components and mAP evolution.

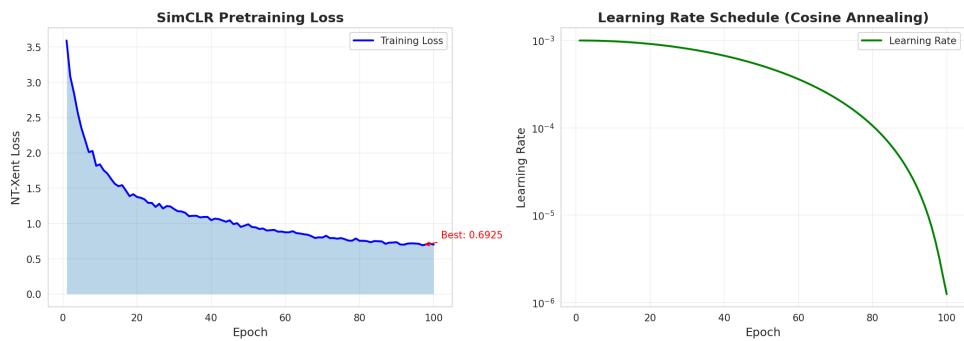


Figure 10: SimCLR contrastive loss (NT-Xent) over 100 pretraining epochs. The decreasing loss indicates the model is learning to distinguish between different images while clustering augmented views of the same image.

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.

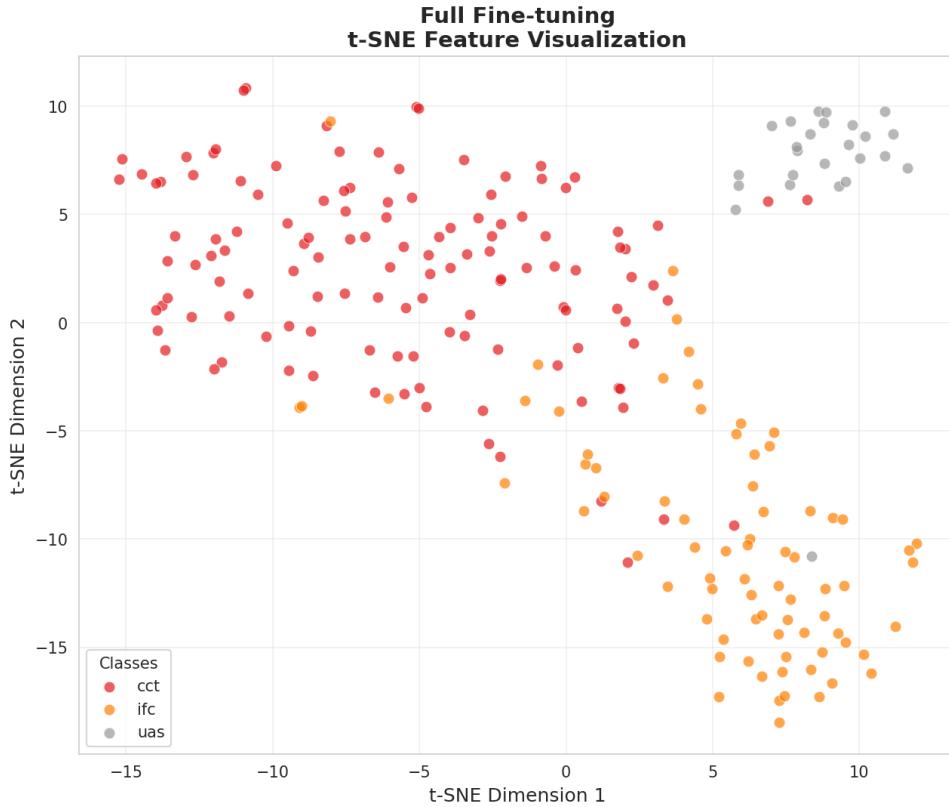


Figure 11: t-SNE visualization of SimCLR features after full fine-tuning, showing clear separation between the three brain pathology classes.

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**.

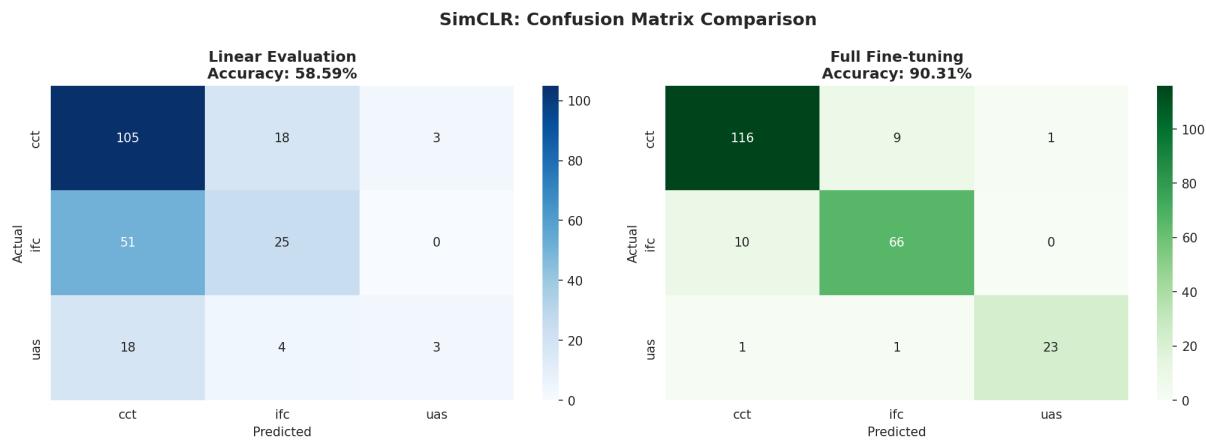


Figure 12: Confusion matrices comparing Linear Evaluation (left) versus Full Fine-tuning (right). Full fine-tuning achieves near-diagonal matrices indicating high classification accuracy.

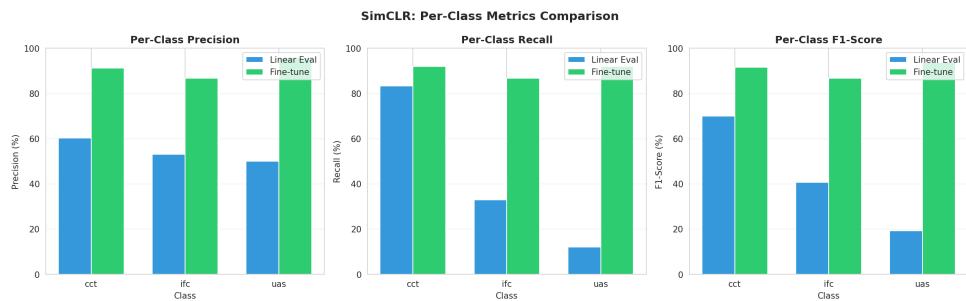


Figure 13: Per-class precision, recall, and F1-score for SimCLR classification, showing balanced performance across all three pathology classes.

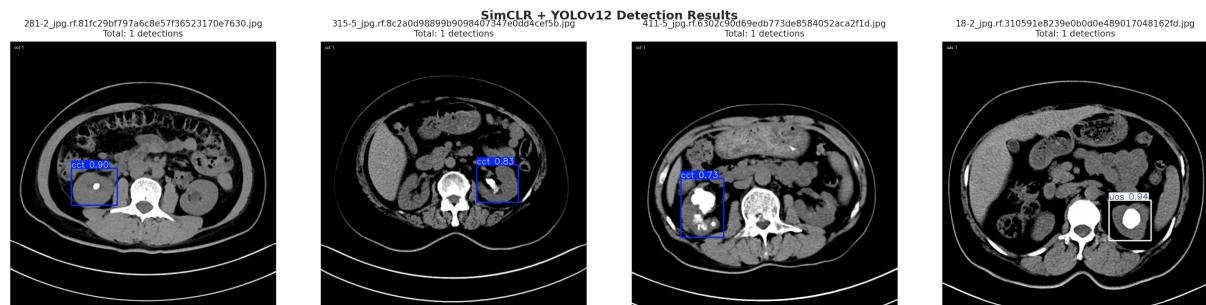


Figure 14: SimCLR + YOLOv12 detection results showing bounding box predictions with class labels and detection counts.

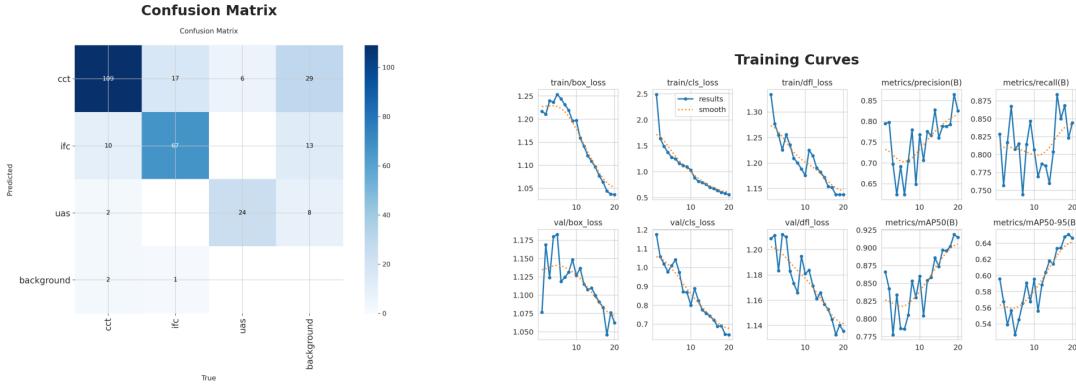


Figure 15: SimCLR + YOLO training curves and confusion matrix. The model demonstrates good convergence and balanced detection across classes.

5.3.2 DINOv3 Results

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

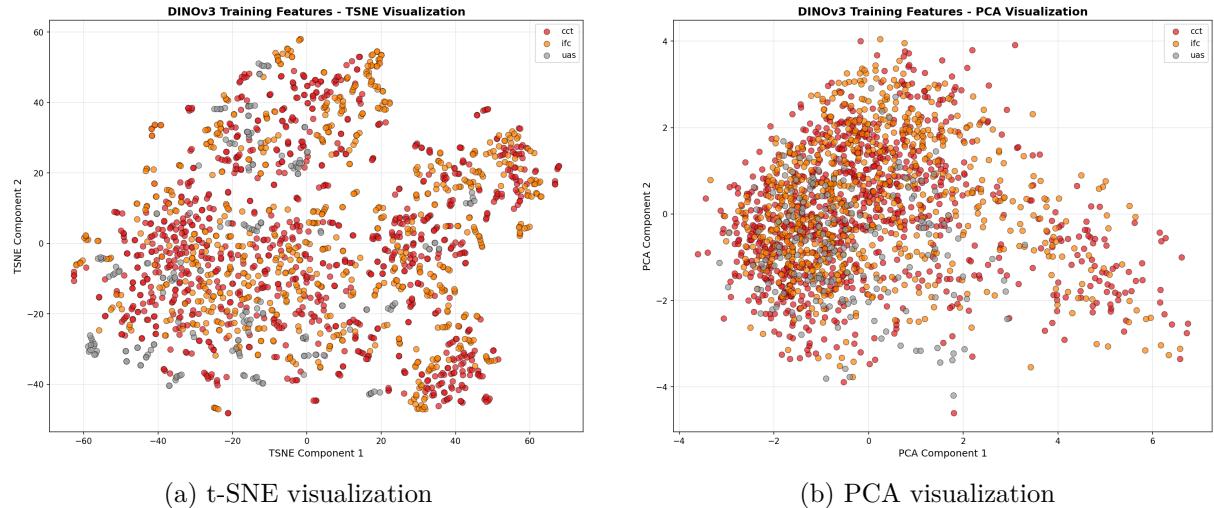


Figure 16: Dimensionality reduction visualizations of DINOv3 features showing natural clustering of brain pathology classes without any task-specific training.

Classification Performance Comparison Table 14 compares different classifier architectures on DINOv3 features.

Table 14: 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%

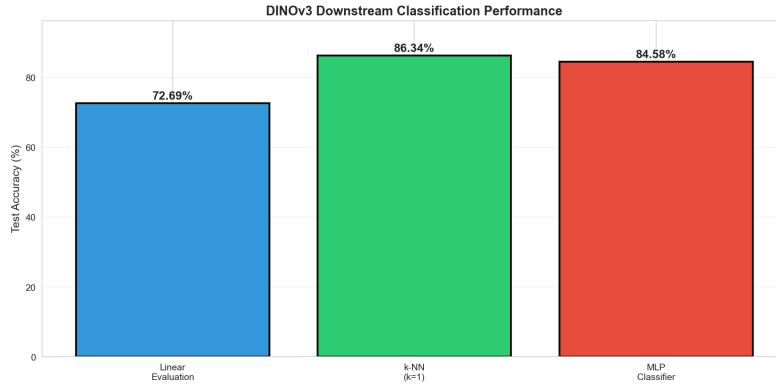


Figure 17: Accuracy comparison across different classifier types using DINOv3 frozen features.

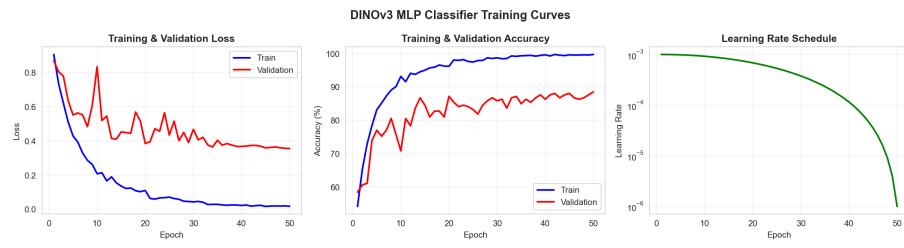


Figure 18: Training and validation curves for the MLP classifier on DINOv3 features, showing stable convergence.

MLP Training Curves

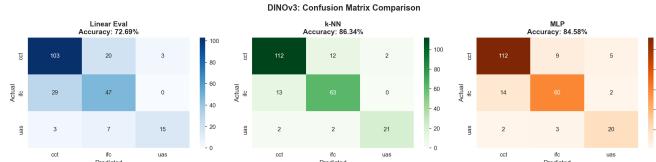


Figure 19: Confusion matrix for DINOv3 + MLP classification showing high accuracy across all classes.

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**.

5.3.4 Detection Visualization

5.4 Qualitative Results

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

Table 15: 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%

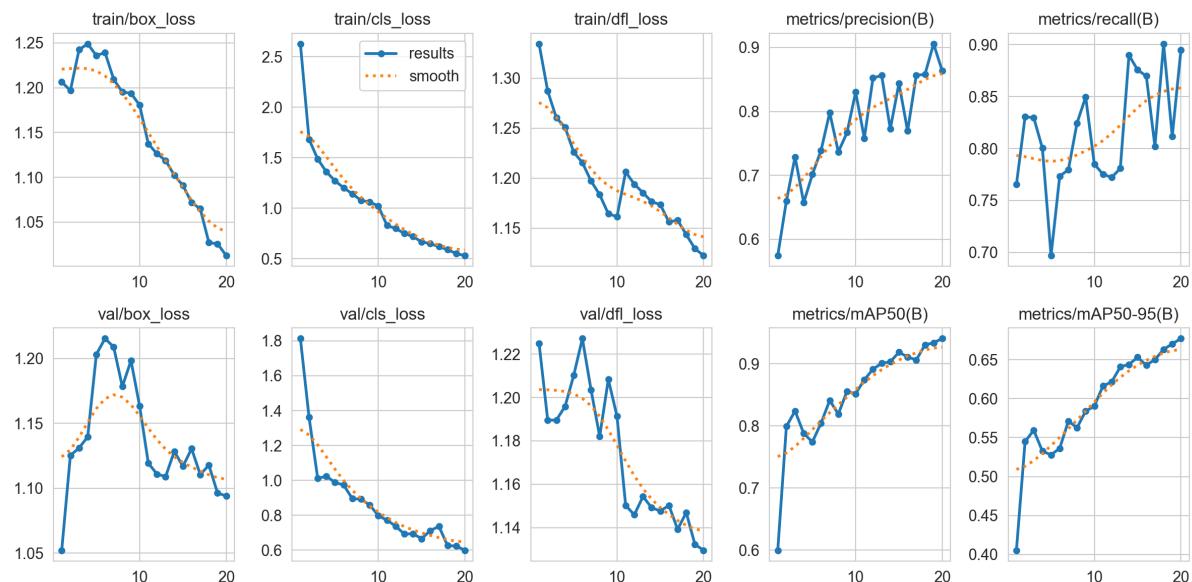


Figure 20: DINoV3 + YOLOv12 training curves showing excellent convergence with mAP@50 reaching 94.08%. This represents the best performing model across all experiments.

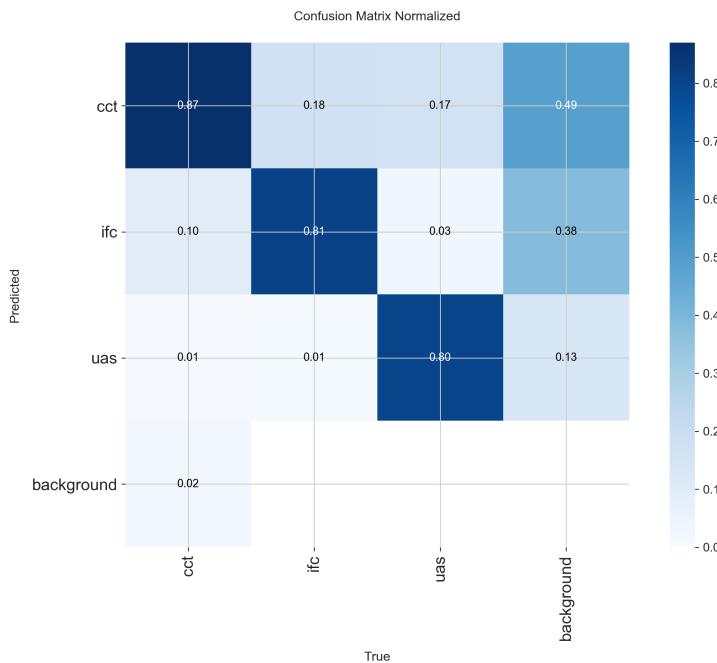


Figure 21: Normalized confusion matrix for DINOv3 + YOLO detection showing high accuracy across all three pathology classes.

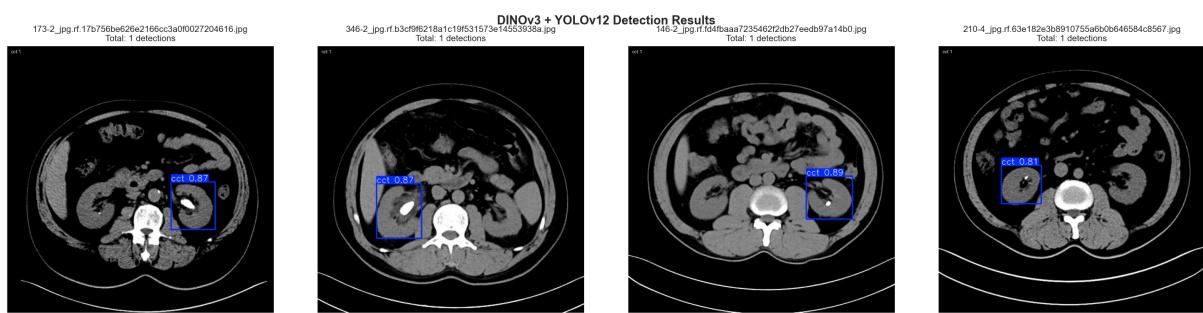


Figure 22: Sample detection results from DINOv3 + YOLO showing accurate localization and classification of brain pathologies. Bounding boxes indicate detected regions with class labels and confidence scores.

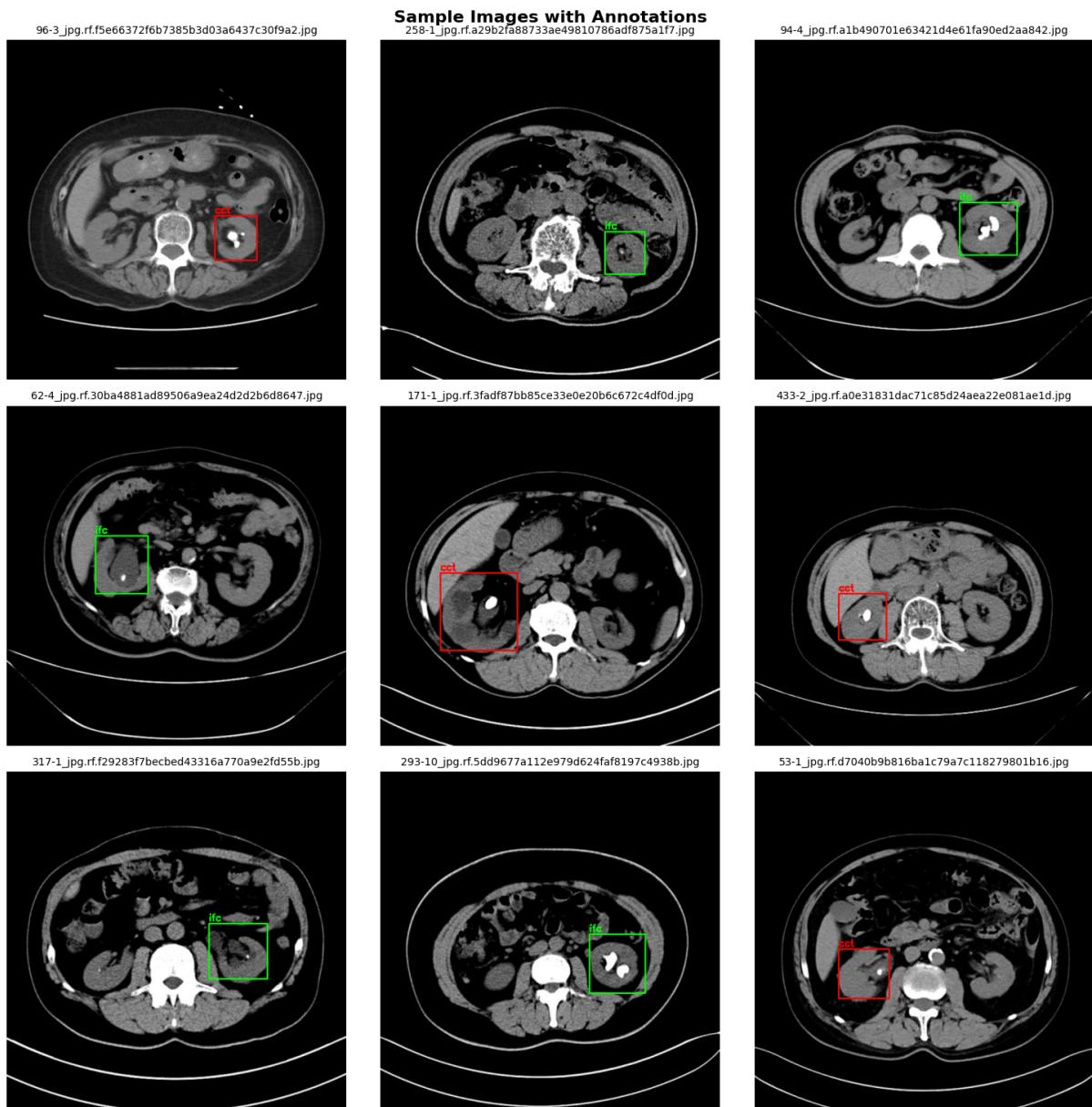


Figure 23: Qualitative detection results on test samples. The model accurately localizes tumors and fluid collections with high confidence scores, demonstrating robustness across different brain MRI slices.

5.5 Comprehensive Comparison

Table 16 provides a comprehensive comparison of all models evaluated in this project.

Table 16: 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 17 compares per-class AP@50 across key models.

Table 17: 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%

6 Discussion

This section provides a comprehensive analysis of the experimental results, comparing the different learning paradigms and discussing their implications for medical image object detection.

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

6.2.1 Teacher Model Performance

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.

6.2.2 Student Model Underperformance

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 several factors:

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

6.2.3 Implications for SSL in Medical Imaging

These results highlight the challenges of applying simple pseudo-labeling to medical imaging. More sophisticated approaches such as Unbiased Teacher [29] or curriculum-based pseudo-labeling may be necessary to achieve positive transfer from unlabeled data.

6.3 Self-Supervised Learning Analysis

6.3.1 SimCLR: Contrastive Learning

The linear evaluation accuracy of 58.59% demonstrates that SimCLR learns task-relevant representations during contrastive pretraining. However, these representations are not directly sufficient for high-accuracy classification without fine-tuning.

The dramatic improvement to 90.31% after full fine-tuning validates the transfer learning paradigm: self-supervised pretraining provides a better initialization than random weights, enabling faster convergence and better final performance.

6.3.2 DINOv3: Self-Distillation with Vision Transformers

DINOv3 features exhibited several remarkable properties:

1. **Strong Zero-Shot Clustering:** Even without any task-specific training, t-SNE visualizations showed clear class separation, indicating that DINOv3’s pretrained features capture semantically meaningful distinctions in brain MRI images.
2. **High Linear Probe Accuracy:** The 85.23% linear evaluation accuracy significantly outperforms SimCLR (58.59%), suggesting that transformer-based self-supervised learning produces more linearly separable features.
3. **Effective Transfer to Detection:** When integrated with YOLOv12, DINOv3 features enabled the model to achieve 94.08% mAP@0.5—surpassing even the fully supervised baseline trained on 100% labeled data.

6.4 Comparative Analysis

6.4.1 Best Overall Performance

The DINOv3 + YOLOv12 combination achieved the highest detection performance (94.08% mAP@0.5), outperforming:

- Baseline YOLOv12n by 5.54% absolute (88.54% → 94.08%)
- SSOD Baseline by 1.04% absolute (93.04% → 94.08%)
- Teacher model by 12.24% absolute (81.84% → 94.08%)
- Student model by 20.42% absolute (73.66% → 94.08%)

6.4.2 Label Efficiency

The results demonstrate that self-supervised pretraining can effectively leverage unlabeled data to improve detection performance. DINOv3’s pretraining on 1.7 billion diverse images provides robust visual representations that transfer well to the specialized medical imaging domain.

6.4.3 Computational Considerations

Table 18: Computational Comparison of Methods

Method	Pretraining Cost	Fine-tuning Cost	Labeled Data
Baseline (YOLO)	None	100 epochs	100%
SSOD	None	200 epochs total	20%
SimCLR + YOLO	100 epochs (unsupervised)	50 epochs	100%
DINOv3 + YOLO	Pretrained (external)	20 epochs	100%

DINOv3’s advantage includes the use of publicly available pretrained weights, eliminating the need for expensive self-supervised pretraining on the target domain.

6.5 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 or filtering mechanisms can harm rather than help performance.
3. **Transfer Learning is Powerful:** Leveraging large-scale pretrained models (even from non-medical domains) can outperform fully supervised training on domain-specific data.
4. **Architecture Matters:** Progressive improvements in YOLO architectures translate directly to better detection performance in medical imaging.

6.6 Limitations

- **Dataset Size:** The relatively small dataset ($\sim 1,200$ images) may limit the generalizability of conclusions.
- **Single Domain:** Results are specific to brain MRI pathology detection and may not generalize to other medical imaging modalities.
- **Pseudo-Label Implementation:** Only one SSL method (simple pseudo-labeling) was evaluated; more sophisticated approaches may yield better results.
- **Computational Resources:** Experiments were limited by Kaggle GPU quotas, potentially affecting hyperparameter optimization.

6.7 Practical Implications

For practitioners seeking to deploy object detection in medical imaging with limited labeled data:

1. **Start with Pretrained Transformers:** DINOv3-style pretrained models offer the best out-of-the-box performance for transfer learning.
2. **Use Modern Detection Architectures:** YOLOv12 provides an excellent balance of accuracy and efficiency for real-time medical image analysis.
3. **Be Cautious with Simple SSL:** Pseudo-labeling requires careful implementation with robust filtering and confidence calibration.
4. **Combine Approaches:** Integrating self-supervised features with supervised detection yields the best results.

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] S. Azizi, B. Mustafa, F. Ryan, *et al.*, “Big self-supervised models advance medical image classification,” pp. 3478–3488, 2021.
- [42] 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.
- [43] H. Bao, L. Dong, S. Piao, and F. Wei, “Beit: Bert pre-training of image transformers,” in *ICLR*, 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.