



SAPIENZA
UNIVERSITÀ DI ROMA

Sviluppo di una pipeline di Deep Learning per la diagnosi di noduli tiroidei in ecografia: confronto tra architetture CNN e Vision Transformers

Facoltà di Ingegneria dell'informazione, informatica e statistica
Corso di Laurea in Informatica

Candidato

Alessandro Catania
Matricola 1916996

Relatore

Prof. Cinque Luigi

Correlatore

Prof. Fagioli Alessio

Anno Accademico 2024/2025

Sviluppo di una pipeline di Deep Learning per la diagnosi di noduli tiroidei in ecografia: confronto tra architetture CNN e Vision Transformers

Tesi di Laurea. Sapienza – Università di Roma

© 2025 Alessandro Catania. Tutti i diritti riservati

Questa tesi è stata composta con L^AT_EX e la classe Sapthesis.

Email dell'autore: catania.1916996@studenti.uniroma1.it

Sommario

La diagnosi dei noduli tiroidei rappresenta una sfida clinica significativa a causa dell'elevata prevalenza della patologia e della complessità interpretativa delle immagini ecografiche, spesso affette da rumore e variabilità operatore-dipendente. Sebbene il Deep Learning abbia dimostrato notevoli capacità nel supporto alla diagnosi medica (*Computer-Aided Diagnosis*, CAD), la letteratura recente è divisa sull'efficacia relativa delle architetture classiche basate su Convoluzioni (CNN) rispetto ai più recenti Vision Transformers (ViT) e Foundation Models in contesti a scarsità di dati o alta rumorosità.

Il presente lavoro di tesi propone una pipeline completa a due stadi per l'analisi automatica dei noduli tiroidei. La prima fase, dedicata alla *Object Detection*, confronta approcci real-time basati su meccanismi di attenzione (YOLOv12) con architetture a due stadi e basate su Transformer (DINO-DETR, Faster R-CNN). La seconda fase, focalizzata sulla classificazione binaria (benigno/maligno), valuta la capacità di generalizzazione di modelli pre-addestrati su larga scala come DINOv3 rispetto a reti efficienti convenzionali (EfficientNetV2, ConvNeXtv2).

Lo studio è stato condotto su un dataset aggregato ed eterogeneo di oltre 7.000 noduli, sottoposto a una rigorosa procedura di pulizia tramite *perceptual hashing* e migliorato mediante tecniche di *image enhancement* (CLAHE, Sharpening). I risultati sperimentali mirano a quantificare se l'introduzione del *Self-Supervised Learning* e dei meccanismi di *Attention* globale offre vantaggi tangibili rispetto alle CNN tradizionali nella distinzione di pattern morfologici sottili, contribuendo a definire lo stato dell'arte per la diagnostica tiroidea assistita.

Indice

1	Introduzione	1
1.1	Motivazione e Contesto Clinico	1
1.2	Evoluzione del Deep Learning: dalle CNN ai Transformers	2
1.3	Obiettivi perseguiti nello studio	2
1.4	Struttura dell'Elaborato	3
2	Background Teorico	4
2.1	Il Dominio Medico: Ecografia e Diagnostica Tiroidea	4
2.1.1	Principi dell'Ecografia B-Mode	4
2.1.2	Il Sistema K-TIRADS	4
2.1.3	Le Sfide dell'Analisi Automatica	6
2.2	L'Approccio Tradizionale: Radiomica e Feature Hand-Crafted	6
2.2.1	Feature Estratte	6
2.3	Evoluzione del Deep Learning: Le CNN	7
2.3.1	EfficientNetV2: Lo Stato dell'Arte delle CNN	7
2.3.2	ConvNeXt V2: Il Ponte tra CNN e Transformer	8
2.4	Vision Transformers e Foundation Models	8
2.4.1	Self-Attention Mechanism	8
2.4.2	DINOv3: Self-Supervised Learning	8
2.5	Architetture per Object Detection	9
2.5.1	Two-Stage Detectors: Faster R-CNN	9
2.5.2	One-Stage Detectors: YOLOv12	10
2.5.3	Transformer-based Detectors: DINO-DETR	10
2.6	Stato dell'Arte: Soluzioni Commerciali e Nuove Frontiere	10
2.6.1	Soluzioni Commerciali Approvate FDA	11
2.6.2	Frontiere della Ricerca	11
2.7	Metriche di Valutazione	11
3	Materiali e Dataset	12
3.1	Descrizione e Fonti dei Dati	12
3.2	Pulizia dei Dati e Rimozione Duplicati	12
3.2.1	Algoritmo di Perceptual Hashing	13
3.3	Analisi Statistica del Dataset	14
3.3.1	Distribuzione delle Classi	14
3.3.2	Analisi Dimensionale e Morfometrica	14
3.4	Image Enhancement e Pre-processing	16
3.4.1	CLAHE e Sharpening	16
3.5	Preparazione Specifica per i Task	16
3.5.1	Input per Detection (Pre-processing Differenziato)	16
3.5.2	Input per Classification (DINOv3 / CNN)	17

3.6	Organizzazione degli Esperimenti	17
4	Metodologia Sperimentale	18
4.1	Ambiente di Sviluppo	18
4.2	Baseline Method: SVM e Radiomic	18
4.3	Fase 1: Configurazione Object Detection	19
4.3.1	YOLOv12 (Ultralytics)	19
4.3.2	DINO-DETR e Faster R-CNN (Detrex)	19
4.4	Fase 2: Configurazione Classificazione	19
4.4.1	Strategia di Training (Fine-Tuning)	20
4.4.2	Iperparametri e Loss Function	20
4.5	Data Augmentation	21
4.6	Protocollo di Riproducibilità	21
5	Risultati e Discussione	23
5.1	Analisi Baseline: Machine Learning Tradizionale (SVM)	23
5.1.1	Risultati Quantitativi	23
5.2	Fase 1: Risultati Object Detection	24
5.2.1	YOLOv12: Analisi Dimensionale (Small vs Medium vs Large)	24
5.2.2	DINO-DETR: Ablation Study sui Backbone	25
5.2.3	Baseline Detection: Faster R-CNN	26
5.3	Fase 2: Risultati Classificazione (Benigno vs Maligno)	26
5.3.1	Confronto Prestazioni Globali	26
5.3.2	Analisi delle Soglie	27
5.4	Analisi Qualitativa: Explainability e Integrazione TI-RADS	27
5.4.1	Confronto Architettonicale: CNN vs Transformer	27
5.5	Analisi dei Casi Critici (Failure Analysis)	29
5.5.1	Falsi Negativi (Maligni classificati come Benigni)	29
5.5.2	Falsi Positivi (Benigni classificati come Maligni)	29
5.5.3	Metodologia: Dai Logits alla Probabilità di Rischio	29
5.5.4	Riconoscimento Implicito delle Feature	30
5.6	Discussione e Integrazione Clinica	31
5.6.1	Proposta di Integrazione con il Workflow TI-RADS	31
5.6.2	Confronto con lo Stato dell'Arte	31
6	Conclusioni e Sviluppi Futuri	32
6.1	Sintesi dei Risultati	32
6.2	Limitazioni dello Studio	33
6.3	Sviluppi Futuri	33
A	Implementazione del Prototipo Demo	34
A.1	Flusso di Utilizzo dell'Interfaccia	34
A.1.1	Pre-processing Interattivo	35
A.1.2	Risultato dell'Inferenza	36
Bibliografia		37
Bibliografia		37

Capitolo 1

Introduzione

La diagnostica per immagini ricopre un ruolo centrale nella medicina moderna, fornendo strumenti non invasivi per l'identificazione precoce e la caratterizzazione di patologie oncologiche. Tra le diverse modalità di imaging, l'ecografia (ultrasuoni) rappresenta il *gold standard* per l'esame della tiroide, una ghiandola endocrina soggetta a un'elevata incidenza di formazioni nodulari.

Il presente lavoro di tesi si inserisce nel contesto dell'Ingegneria Biomedica e dell'Intelligenza Artificiale, proponendo lo sviluppo e l'analisi comparativa di sistemi avanzati di *Computer-Aided Diagnosis* (CAD) basati su Deep Learning, con l'obiettivo di supportare il radiologo nella complessa task di rilevamento e classificazione dei noduli tiroidei.

1.1 Motivazione e Contesto Clinico

I noduli tiroidei sono formazioni estremamente comuni nella popolazione adulta: studi epidemiologici stimano che fino al 50-60% degli adulti sani possa presentare noduli rilevabili ecograficamente. Tuttavia, solo una piccola frazione di questi (circa il 5-10%) risulta essere maligna. La sfida clinica risiede, dunque, nella capacità di discriminare accuratamente i noduli sospetti, che richiedono un accertamento invasivo tramite agoaspirato (FNA - *Fine Needle Aspiration*), da quelli benigni che necessitano solo di monitoraggio.

L'ecografia è l'esame di prima istanza grazie alla sua non invasività, assenza di radiazioni ionizzanti e basso costo. Tuttavia, essa presenta limitazioni intrinseche significative:

- 1. Bassa qualità dell'immagine:** Le immagini ecografiche sono affette da *speckle noise* (rumore granulare), basso contrasto e artefatti acustici che rendono difficile la distinzione dei margini del nodulo.
- 2. Soggettività:** L'interpretazione è fortemente "operatore-dipendente". La classificazione del rischio (spesso basata su protocolli come il TI-RADS) varia significativamente in base all'esperienza del radiologo.

Sebbene l'interpretazione umana resti soggettiva, il panorama clinico sta evolvendo rapidamente con l'introduzione delle prime soluzioni commerciali approvate dagli enti regolatori. Piattaforme come **Koios DST™ Thyroid** e **See-Mode**, recentemente approvate dalla FDA (*Food and Drug Administration*), stanno iniziando a supportare i radiologi con analisi morfologiche automatizzate. Tuttavia, trattandosi di sistemi proprietari "chiusi" (*Black Box*), la ricerca accademica aperta resta cruciale

per comprendere, validare e migliorare le architetture sottostanti, esplorando nuove frontiere come i Large Language Models applicati alla diagnostica (es. ThyGPT).

1.2 Evoluzione del Deep Learning: dalle CNN ai Transformers

Negli ultimi dieci anni, la Computer Vision applicata al *medical imaging* è stata dominata dalle Reti Neurali Convoluzionali (CNN). Architetture come ResNet o EfficientNetv2 hanno raggiunto prestazioni sovraumane in molti task, sfruttando la capacità delle convoluzioni di estrarre caratteristiche locali (bordi, texture) in modo gerarchico. Tuttavia, le CNN possiedono un "campo recettivo" limitato: faticano a catturare relazioni a lungo raggio tra parti distanti dell'immagine, un aspetto che può essere cruciale quando la diagnosi dipende dal contesto globale del tessuto o dalla relazione del nodulo con le strutture adiacenti.

Recentemente, l'introduzione dei **Vision Transformers (ViT)** ha rivoluzionato il settore. Grazie al meccanismo di *Self-Attention*, queste reti sono in grado di modellare dipendenze globali sull'intera immagine fin dai primi strati. Inoltre, l'avvento dei **Foundation Models** (modelli addestrati su quantità massive di dati con tecniche auto-supervisionate, come la serie DINO di Meta) promette di superare le limitazioni dei dataset medici, che sono spesso di dimensioni ridotte rispetto ai dataset generici.

Sorge quindi una domanda di ricerca fondamentale, che guida questa tesi: *Le moderne architetture basate su Transformer e Attention (come DINOv3 o YOLOv12) offrono un reale vantaggio prestazionale rispetto alle consolidate CNN nel dominio specifico, rumoroso e visivamente complesso dell'ecografia tiroidea?*

1.3 Obiettivi perseguiti nello studio

L'obiettivo principale di questo lavoro è progettare, addestrare e validare una pipeline automatizzata a due stadi per l'analisi dei noduli tiroidei, confrontando l'efficacia di paradigmi architetturali diversi.

Gli obiettivi specifici sono:

1. **Creazione di un Dataset Robusto:** Aggregazione e pulizia di fonti dati eterogenee (oltre 7.000 noduli) con tecniche avanzate di *deduplication* e *image enhancement* per mitigare il rumore ecografico.
2. **Analisi della Fase di Detection:** Confrontare l'accuratezza nella localizzazione dei noduli tra rilevatori *Real-Time* di nuova generazione (YOLOv12 con moduli di attention), rilevatori *Transformer-based* (DINO-DETR) e rilevatori classici *Two-Stage* (Faster R-CNN).
3. **Analisi della Fase di Classificazione:** Valutare se l'utilizzo di un Foundation Model come DINOv3 (con fine-tuning completo) permetta di ottenere una classificazione Benigno/Maligno superiore rispetto a CNN moderne (EfficientNetv2, ConvNeXt v2), specialmente in presenza di *crop* che includono il contesto perilesionale.
4. **Valutazione Critica:** Analizzare non solo le metriche quantitative, ma anche il comportamento qualitativo dei modelli per comprendere quali caratteristiche morfologiche vengano privilegiate dalle diverse architetture.

1.4 Struttura dell'Elaborato

La tesi è organizzata come segue:

- Il **Capitolo 2** fornisce il background teorico, descrivendo le sfide dell'imaging ecografico e i fondamenti delle architetture di Deep Learning utilizzate (CNN, Transformers, Object Detection, Self-Supervised Learning).
- Il **Capitolo 3** descrive in dettaglio il dataset, la metodologia di raccolta, le tecniche di *hashing* per la rimozione dei duplicati e la pipeline di pre-processing (CLAHE, Sharpening).
- Il **Capitolo 4** illustra la metodologia sperimentale, dettagliando le configurazioni dei modelli (YOLOv12, DINO-DETR, DINov3, ecc.), le strategie di *data augmentation* e i protocolli di addestramento.
- Il **Capitolo 5** presenta i risultati sperimentali, con un confronto quantitativo (metriche di performance) e qualitativo tra i diversi approcci.
- Il **Capitolo 6** conclude il lavoro, discutendo le implicazioni dei risultati ottenuti e suggerendo possibili sviluppi futuri.

Capitolo 2

Background Teorico

In questo capitolo vengono analizzati i fondamenti teorici e tecnologici che costituiscono la base del presente lavoro di tesi. La trattazione è strutturata per guidare il lettore dal contesto clinico alle metodologie ingegneristiche: si parte dall'analisi dell'ecografia tiroidea e del sistema TI-RADS, si passa per l'approccio tradizionale della radiomica, fino ad arrivare allo stato dell'arte del Deep Learning (con un focus specifico sul confronto CNN vs Vision Transformers) e alle soluzioni commerciali attualmente disponibili.

2.1 Il Dominio Medico: Ecografia e Diagnostica Tiroidea

La tiroide è una ghiandola endocrina situata alla base del collo, fondamentale per la regolazione del metabolismo. La patologia nodulare tiroidea è estremamente diffusa, con una prevalenza che aumenta con l'età e varia a seconda dell'apporto iodico della popolazione.

2.1.1 Principi dell'Ecografia B-Mode

L'ecografia (o ultrasonografia) rappresenta la metodica di *imaging* di prima scelta per lo studio della tiroide. Essa si basa sull'emissione di onde sonore ad alta frequenza (tra 7 e 15 MHz per le sonde lineari superficiali) che, attraversando i tessuti biologici, vengono parzialmente riflesse dalle interfacce acustiche. Il segnale di ritorno viene elaborato e visualizzato in modalità **B-Mode (Brightness Mode)**, dove l'intensità dell'eco riflesso determina la luminosità del pixel corrispondente:

- **Anecogeno (Nero):** Strutture liquide (es. cisti, vasi) che non riflettono il suono.
- **Iperecogeno (Bianco):** Strutture dense o calcificazioni che riflettono fortemente.
- **Isoecogeno/Ipoecogeno (Grigio):** Tessuti molli con diversa densità cellulare rispetto al parenchima sano.

2.1.2 Il Sistema K-TIRADS

Per ridurre la soggettività nella refertazione ecografica e standardizzare la gestione dei pazienti, la Korean Society of Thyroid Radiology ha sviluppato il sistema

K-TIRADS (Korean Thyroid Imaging Reporting and Data System). Questo protocollo classifica i noduli tiroidei non con un semplice punteggio numerico cumulativo, ma secondo un modello basato su pattern ecografici: composizione, ecogenicità e la presenza di caratteristiche sospette, che portano a diverse categorie di rischio.

Criteri principali:

- **Composizione:** suddivisa in solido, prevalentemente solido ($\leq 50\%$ di componente cistica), prevalentemente cistico o cistico puro.
- **Ecogenicità:** definita rispetto al parenchima tiroideo normale e ai muscoli anteriori del collo: marcatamente ipoecogeno, lievemente ipoecogeno, isoecogeno o iperecogeno.
- **Orientamento (forma):** valutato nel piano trasversale; la forma “non parallela” (*taller-than-wide*) è considerata sospetta.
- **Margini:** lisci vs irregolari (spiculati o microlobulati).
- **Foci ecogeni:** include microcalcificazioni, macrocalcificazioni, rim calcificazioni o foci intracisticci con effetto “comet-tail”.

Categorie di rischio:

- **K-TIRADS 2 (benigno):** noduli spongiformi, cistici puri o con foci intracisticci benigni.
- **K-TIRADS 3 (basso sospetto):** noduli parzialmente cistici o iso-/iperecogeni senza caratteristiche sospette.
- **K-TIRADS 4 (sospetto intermedio):** noduli solidi ipoecogeni senza caratteristiche sospette oppure noduli cistici o isoecogeni con almeno una caratteristica sospetta.
- **K-TIRADS 5 (sospetto elevato):** noduli solidi ipoecogeni con almeno una delle caratteristiche sospette: microcalcificazioni, margini irregolari, orientamento non parallelo.

Note aggiuntive: Il sistema definisce anche soglie di dimensione per la biopsia (FNA) che dipendono dalla categoria di rischio, riducendo i prelievi non necessari pur mantenendo sensibilità sufficiente.

È importante notare che anche gli algoritmi di intelligenza artificiale devono imparare a riconoscere queste caratteristiche ecografiche (composizione, ecogenicità, margini, foci) per poter stratificare correttamente il rischio secondo K-TIRADS.

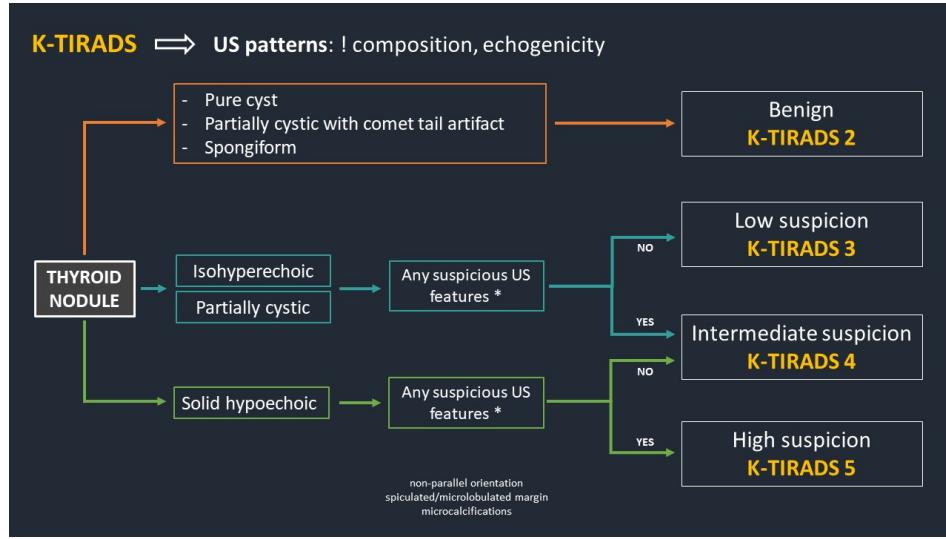


Figura 2.1. Schema di classificazione K-TIRADS: i noduli tiroidei vengono stratificati in categorie di rischio (benigno-basso, intermedio, alto) in base a composizione ecografica, ecogenicità e alla presenza di caratteristiche sospette (es. microcalcificazioni, margini irregolari, orientamento «taller-than-wide»).

2.1.3 Le Sfide dell'Analisi Automatica

Nonostante la standardizzazione, l'interpretazione rimane complessa a causa di artefatti intrinseci all'ecografia:

- **Speckle Noise:** Un rumore granulare moltiplicativo causato dall'interferenza costruttiva e distruttiva delle onde ultrasonore, che degrada la risoluzione dei bordi e riduce il contrasto.
- **Ombre Acustiche:** Zone scure posteriori a calcificazioni che oscurano i tessuti sottostanti.
- **Variabilità Strumentale:** Differenze significative nella qualità dell'immagine a seconda del produttore dell'ecografo.

2.2 L'Approccio Tradizionale: Radiomica e Feature Hand-Crafted

Prima dell'avvento del Deep Learning *end-to-end*, l'analisi computerizzata (CAD) si basava sulla **Radiomica**, ovvero l'estrazione matematica di feature descrittive dall'immagine, seguita da classificatori classici. In questa tesi, questo approccio è utilizzato come *baseline* di confronto.

2.2.1 Feature Estratte

Le feature radiomiche cercano di quantificare matematicamente i concetti del TI-RADS:

- **Morfologia (Forma):** La metrica più rilevante è l'**Aspect Ratio**.

$$\text{Aspect Ratio} = \frac{\text{Altezza}}{\text{Larghezza}} \quad (2.1)$$

Un valore > 1 corrisponde al criterio "più alto che largo", forte indicatore di malignità (crescita centrifuga del tumore contro i piani tissutali).

- **Texture (GLCM):** Si utilizzano descrittori statistici del secondo ordine basati sulla **Grey Level Co-occurrence Matrix (GLCM)**. Metriche come *Contrasto*, *Omogeneità* ed *Energia* misurano la variazione spaziale dei livelli di grigio.
- **Statistica del Primo Ordine:** Metriche basate sull'istogramma dei pixel (es. *Skewness* e *Kurtosis*), utili per descrivere la distribuzione globale dell'ecogenicità.

Queste feature vengono tipicamente utilizzate come input per classificatori come le **Support Vector Machines (SVM)**.

2.3 Evoluzione del Deep Learning: Le CNN

Il Deep Learning supera i limiti della radiomicia apprendendo automaticamente le feature ottimali direttamente dai dati (*Representation Learning*).

2.3.1 EfficientNetV2: Lo Stato dell'Arte delle CNN

EfficientNetV2 rappresenta l'apice dell'evoluzione delle CNN tradizionali ed è utilizzata in questo studio come *strong baseline* convoluzionale. Rispetto alle versioni precedenti, introduce innovazioni mirate alla velocità di addestramento e all'efficienza parametrica:

1. **Fused-MBConv:** Nei primi stadi della rete, le costose convoluzioni *depthwise* 3×3 e *pointwise* 1×1 vengono fuse in un'unica convoluzione standard 3×3 . Questo riduce l'overhead di accesso alla memoria, ottimizzando l'uso delle moderne GPU.
2. **Compound Scaling:** La rete scala uniformemente tutte le dimensioni (profondità, larghezza, risoluzione) tramite un coefficiente composto.
3. **Progressive Learning:** Durante il training, la dimensione delle immagini aumenta progressivamente, adattando dinamicamente la regolarizzazione per prevenire l'overfitting.

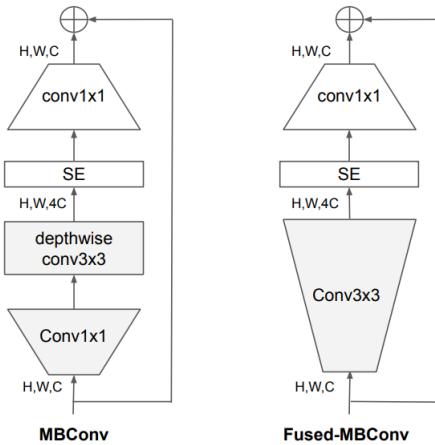


Figura 2.2. Confronto tra il blocco MBConv standard e il Fused-MBConv introdotto in EfficientNetV2 per migliorare l’efficienza hardware.

2.3.2 ConvNeXt V2: Il Ponte tra CNN e Transformer

ConvNeXtV2 estende l’idea di “modernizzare” le CNN introducendo ulteriori affinamenti architettonici e un addestramento più stabile. Mantiene la struttura gerarchica tipica della famiglia, ma integra blocchi con **Global Response Normalization (GRN)**, che migliora la sensibilità ai pattern globali e stabilizza le attivazioni. Inoltre adotta **depthwise convolution** con kernel ampi, **Layer Normalization**, attivazioni **GELU** e un design ancora più semplificato e regolare, ottimizzato sia per il training da zero sia per il transfer learning. Il risultato è una CNN più robusta, capace di avvicinarsi ulteriormente alle prestazioni dei Vision Transformers pur conservando l’efficienza delle convoluzioni.

2.4 Vision Transformers e Foundation Models

2.4.1 Self-Attention Mechanism

I Vision Transformers (ViT) abbandonano le convoluzioni locali a favore del meccanismo di **Self-Attention**, che permette a ogni porzione dell’immagine (*patch*) di relazionarsi con tutte le altre contemporaneamente. L’equazione fondamentale è:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (2.2)$$

Questo conferisce alla rete un **campo recettivo globale** nativo, ideale per cogliere la forma complessiva del nodulo e il suo rapporto con il contesto anatomico circostante.

2.4.2 DINOv3: Self-Supervised Learning

DINOv3 (Distillation with No Labels) è un *Foundation Model* basato su Vision Transformer. La sua caratteristica rivoluzionaria è l’addestramento **Self-Supervised (SSL)** su dataset massivi. Il modello utilizza un paradigma "Teacher-Student": due reti con la stessa architettura processano viste diverse (ritagli locali vs globali) della

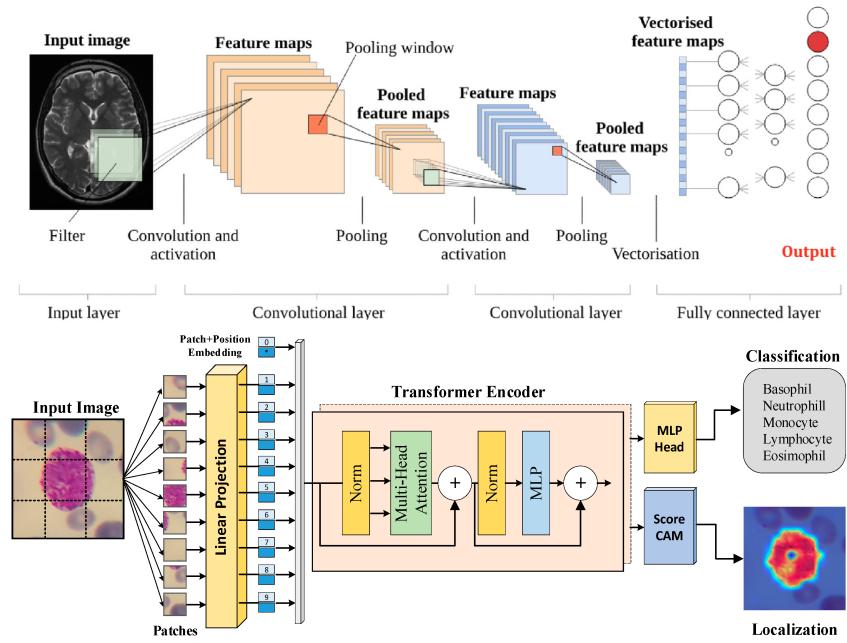


Figura 2.3. Differenza strutturale tra CNN e ViT. Le CNN (sopra) elaborano informazioni locali; i Transformer (sotto) collegano ogni patch dell'immagine con tutte le altre.

stessa immagine. La rete deve imparare a far coincidere le rappresentazioni latenti senza l'uso di etichette manuali. Questo approccio permette a DINOv3 di apprendere feature semantiche estremamente robuste (come la segmentazione implicita degli oggetti) che si trasferiscono efficacemente al dominio medico.

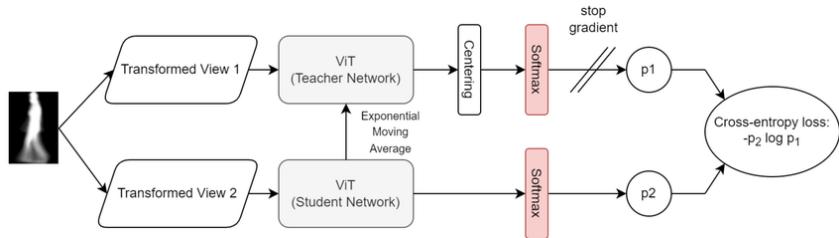


Figura 2.4. Schema del training Self-Supervised in DINO. La rete impara a generare rappresentazioni consistenti a partire da viste diverse della stessa immagine.

2.5 Architetture per Object Detection

Il task di *Object Detection* combina la localizzazione (regressione delle coordinate del bounding box) con la classificazione dell'oggetto.

2.5.1 Two-Stage Detectors: Faster R-CNN

Faster R-CNN è l'archetipo dei rilevatori a due stadi e rappresenta la baseline storica per questo studio. Utilizza una *Region Proposal Network* (RPN) per proporre regioni di interesse, seguita da una testa di classificazione. Sebbene accurato, tende a essere lento e a generare un alto numero di falsi positivi.

2.5.2 One-Stage Detectors: YOLOv12

La famiglia **YOLO** (You Only Look Once) ha rivoluzionato il settore unificando i due stadi in un singolo passaggio. **YOLOv12** rappresenta l'ultima evoluzione (2024/25). Rispetto ai predecessori, introduce la **Anchor-Free Detection** e integra moduli di **Attention** nel backbone per migliorare la focalizzazione sui dettagli, colmando il gap di accuratezza con i modelli a due stadi pur mantenendo velocità *Real-Time*.

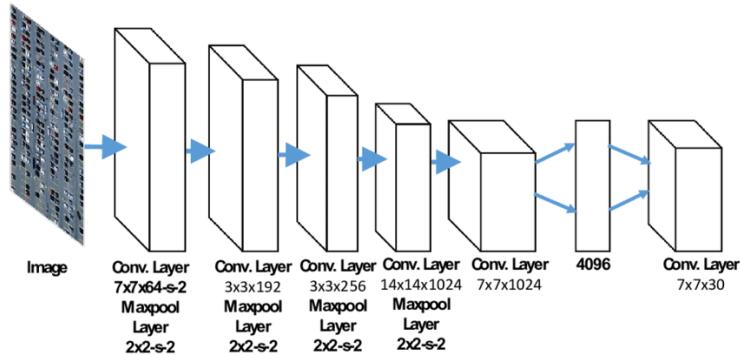


Figura 2.5. Schema generale dell'architettura YOLO. L'immagine viene processata in un unico passaggio attraverso il Backbone, il Neck e la Head.

2.5.3 Transformer-based Detectors: DINO-DETR

DINO (Denoising Inspector for Next-generation Object detection) si basa sull'architettura DETR. Tratta la detection come un problema di predizione di insieme, eliminando la *Non-Maximum Suppression* (NMS) manuale. Utilizza un algoritmo di "matching bipartito" e una tecnica di *Denoising Training* per convergere più rapidamente. È noto per l'estrema precisione geometrica dei box predetti.

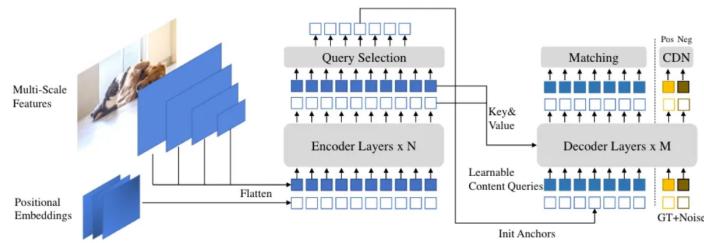


Figura 2.6. Architettura DINO-DETR. Il modello utilizza un Encoder-Decoder Transformer per predire direttamente l'insieme dei box senza anchor.

2.6 Stato dell'Arte: Soluzioni Commerciali e Nuove Frontiere

Il panorama della diagnostica tiroidea assistita sta vedendo l'ingresso delle prime soluzioni certificate, parallelamente alla ricerca avanzata.

2.6.1 Soluzioni Commerciali Approvate FDA

A differenza dei modelli di ricerca, i software commerciali devono superare rigorosi test clinici.

- **Koios DS™ Thyroid:** Software approvato FDA che utilizza il Deep Learning per classificare i noduli secondo le categorie TI-RADS, con l'obiettivo primario di ridurre le biopsie non necessarie (aumentando la specificità).
- **See-Mode:** Soluzione focalizzata sull'analisi automatica dell'intero stack di immagini e sulla generazione di reportistica strutturata.

2.6.2 Frontiere della Ricerca

La ricerca accademica sta esplorando l'uso di **Large Language Models (LLM)** e AI Generativa. Esempi come **ThyGPT** mostrano come i modelli possano non solo classificare, ma generare spiegazioni testuali della diagnosi (*Explainable AI*), una direzione verso cui anche questa tesi si muove attraverso l'analisi delle *heatmap*.

2.7 Metriche di Valutazione

Per validare i modelli, vengono utilizzate metriche standard:

- **IoU (Intersection over Union):** Misura la sovrapposizione tra box predetto e reale.
- **mAP (Mean Average Precision):** Area sotto la curva Precision-Recall per la detection.
- **Sensitivity (Recall):** Capacità di trovare i casi positivi (maligni). Fondamentale per lo screening.
- **AUC-ROC:** Valuta la capacità discriminativa globale del classificatore.

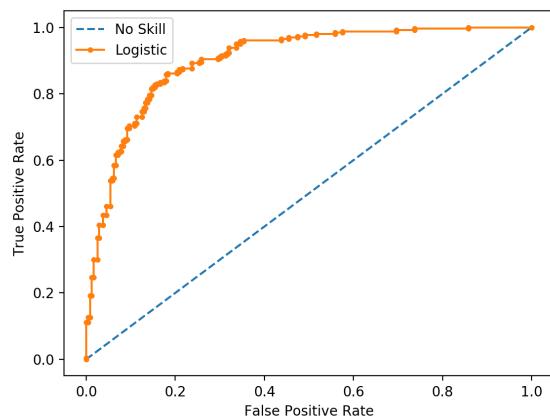


Figura 2.7. Esempio di curva ROC. Un valore di AUC (Area Under Curve) vicino a 1 indica un classificatore ideale.

Capitolo 3

Materiali e Dataset

La qualità dei dati e la robustezza delle procedure di pre-processing sono fattori determinanti per il successo di un sistema basato su Deep Learning, specialmente in ambito medico dove la variabilità inter-paziente, il rumore strumentale e le differenze nei protocolli di acquisizione sono elevati. In questo capitolo viene descritto il dataset eterogeneo costruito per questo studio, dettagliando le fonti, le rigorose procedure di deduplicazione applicate per evitare il fenomeno del *data leakage*, l'analisi statistica della popolazione e la pipeline di *Image Enhancement* sviluppata per mitigare le problematiche tipiche dell'ecografia.

3.1 Descrizione e Fonti dei Dati

Al fine di garantire che i modelli addestrati siano in grado di generalizzare su diverse popolazioni e strumentazioni ecografiche, evitando l'*overfitting* su un singolo centro medico, è stato selezionato un dataset eterogeneo che rappresenta lo stato dell'arte attuale per la diagnostica tiroidea.

La fonte principale, che costituisce il nucleo del training set, è il dataset **TN5000 (Thyroid Nodule 5000)**, una vasta collezione rilasciata recentemente su *Nature Scientific Data*. Questo dataset rappresenta un avanzamento significativo rispetto alle risorse precedenti, fornendo circa 5.000 immagini ecografiche B-mode con annotazioni di alta qualità. A differenza di molti dataset pubblici che si basano solo sulla classificazione visiva (TI-RADS stimato), il TN5000 offre etichette confermate da esame istologico o citologico (biopsia FNA) e maschere di segmentazione precise verificate da radiologi esperti, garantendo un *Ground Truth* estremamente affidabile.

A questo nucleo principale è stato affiancato il dataset **AUITD (Algeria Ultrasound Images Thyroid Dataset)** per introdurre ulteriore variabilità in termini di etnia dei pazienti e tipologia di ecografi utilizzati (es. sonde Toshiba/Samsung con diverse frequenze e impostazioni di guadagno), migliorando la robustezza del modello a variazioni strumentali.

Le annotazioni utilizzate per questo studio consistono in *Bounding Box* (derivate dalle maschere di segmentazione nel caso del TN5000) che delimitano la regione del nodulo e la relativa etichetta di classificazione binaria (Benigno/Maligno).

3.2 Pulizia dei Dati e Rimozione Duplicati

L'aggregazione di dataset pubblici comporta un rischio critico spesso sottovalutato: la duplicazione delle immagini. È frequente che lo stesso caso clinico sia presente in più archivi con nomi file diversi o che vengano salvati frame consecutivi (quasi

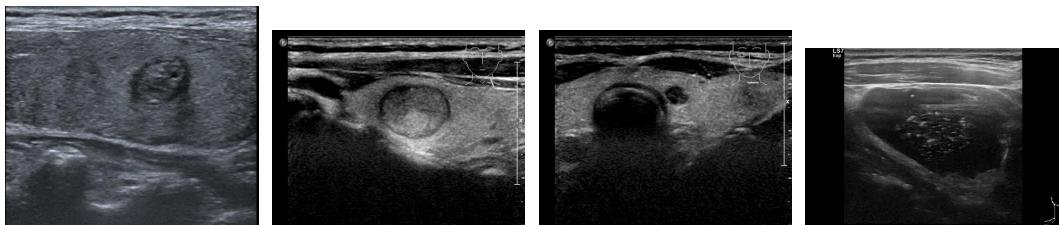


Figura 3.1. Esempi rappresentativi del dataset aggregato. La variabilità nella luminosità, nel contrasto e nella risoluzione riflette l'eterogeneità delle fonti di acquisizione multicentriche.

identici) dello stesso video ecografico. Se non gestita, questa ridondanza porta al fenomeno del **Data Leakage**: se un'immagine duplicata finisce sia nel *Training Set* che nel *Test Set*, il modello otterrà prestazioni falsamente elevate, avendo di fatto "memorizzato" la risposta del test.

Per garantire l'integrità scientifica dello studio, è stata implementata una pipeline automatizzata di deduplica basata su **Perceptual Hashing**.

3.2.1 Algoritmo di Perceptual Hashing

A differenza delle funzioni di hash crittografico (come MD5) che cambiano completamente al variare di un singolo bit, il Perceptual Hashing genera un'impronta digitale ("hash") basata sul contenuto visivo dell'immagine. La procedura attuata è stata la seguente:

1. Calcolo dell'hash per ogni immagine utilizzando l'algoritmo **dhash** (Difference Hash), resistente a lievi ridimensionamenti o variazioni di formato.
2. Confronto tra gli hash tramite **Distanza di Hamming**.
3. Rimozione automatica delle immagini con distanza di Hamming inferiore a una soglia critica, identificando e scartando le copie visivamente identiche.

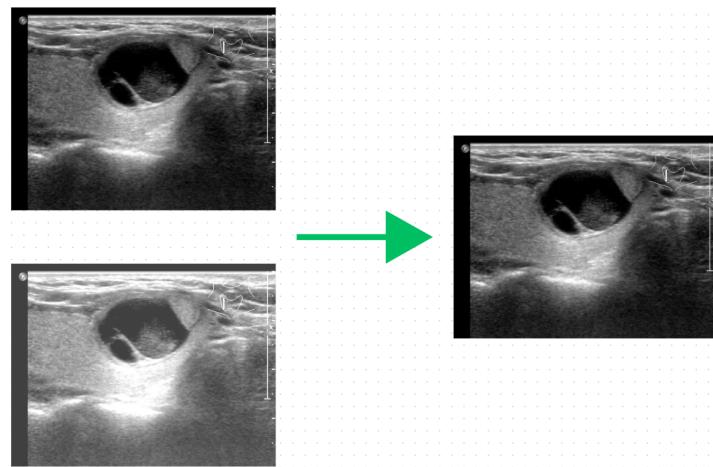


Figura 3.2. Schema della procedura di deduplica tramite Hashing Percettivo. Le immagini con impronta visiva sovrapponibile vengono epurate per mantenere una singola istanza univoca.

3.3 Analisi Statistica del Dataset

A valle della procedura di pulizia, il dataset finale risulta composto da un totale di **7.174 noduli**. Al fine di comprendere le caratteristiche della popolazione in esame e prevenire potenziali bias durante il training, è stata condotta un'analisi statistica descrittiva dettagliata.

3.3.1 Distribuzione delle Classi

La distribuzione delle etichette conferma una prevalenza della classe patologica, situazione atipica per lo screening di popolazione generale ma frequente nei dataset clinici specializzati (dataset terziari):

- **Classe 0 (Benigno):** 2.840 noduli (39.6%)
- **Classe 1 (Maligno):** 4.334 noduli (60.4%)

Nonostante lo sbilanciamento, la classe minoritaria (Benigna) conta quasi 3.000 esemplari, fornendo una base statistica sufficiente per l'apprendimento delle feature discriminative senza la necessità di tecniche aggressive di *oversampling* sintetico.

3.3.2 Analisi Dimensionale e Morfometrica

L'analisi dei diametri dei noduli (espressi in pixel sui bounding box) ha evidenziato differenze morfologiche significative tra le due classi.

- **Statistiche Generali:** Media globale di 200.1 px (Dev. Std: 131.3). I noduli variano da formazioni millimetriche (17 px) a masse massive che occupano l'intero campo visivo (717 px).
- **Noduli Benigni:** Presentano dimensioni mediamente maggiori (Mediana 176.0 px, Media 213.1 px) e una distribuzione molto ampia (*Interquartile Range* IQR = 233.0). Il valore di Kurtosis negativo (-0.43) indica una distribuzione "piatta" (platicurtica): i noduli benigni sono estremamente variegati in dimensione, spaziando uniformemente da piccoli a molto grandi (es. gozzi multinodulari o cisti voluminose).
- **Noduli Maligni:** Tendono a essere più piccoli (Mediana 153.0 px, Media 191.5 px) e più concentrati attorno al valore mediano (IQR più stretto di 130.0). Tuttavia, presentano una distribuzione "appuntita" (leptocurtica, Kurtosis 1.86) con una coda destra pronunciata (Skewness 1.47) e un elevato numero di *outliers* superiori (245 noduli).

Questa analisi suggerisce che le reti neurali dovranno affrontare due sfide opposte: generalizzare su noduli benigni di qualsiasi dimensione (alta varianza) e identificare noduli maligni prevalentemente piccoli, gestendo però correttamente anche i rari casi di carcinomi molto estesi, evitando che vengano scambiati per benigni solo a causa della loro dimensione.

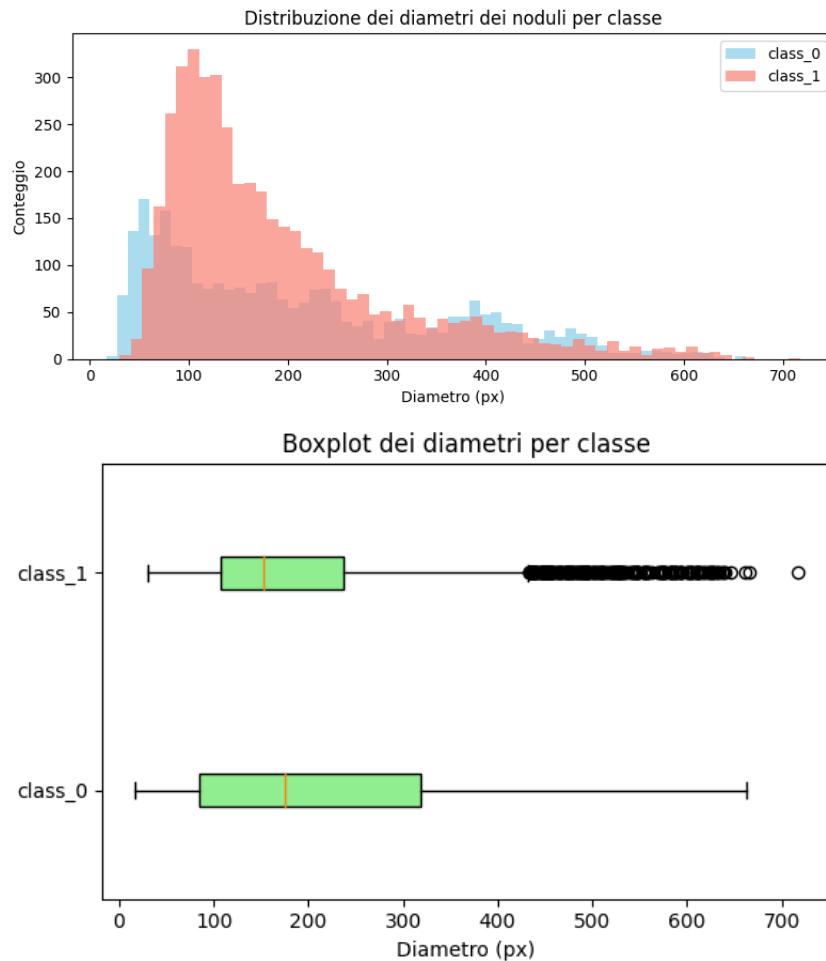


Figura 3.3. Analisi dimensionale dei noduli per classe. In alto: Istogramma della distribuzione dei diametri. In basso: Boxplot che evidenzia la maggiore varianza della classe benigna (class_0) e la presenza di numerosi outliers nella classe maligna.

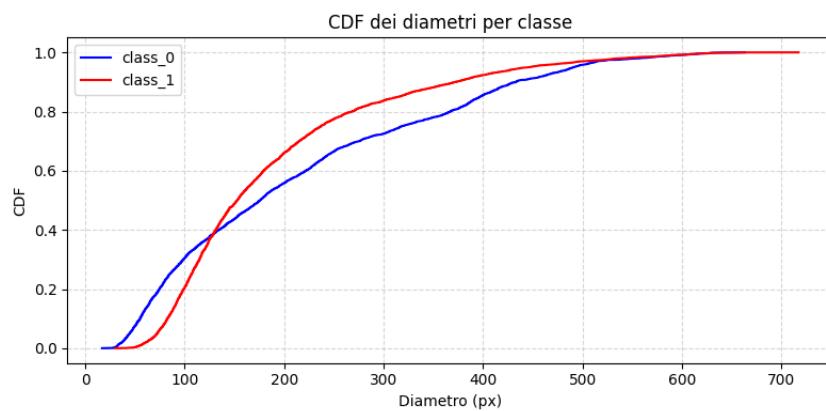


Figura 3.4. Funzione di Distribuzione Cumulativa (CDF) dei diametri. Si nota come la curva dei noduli maligni (rossa) cresca più rapidamente inizialmente, indicando una prevalenza di noduli piccoli.

3.4 Image Enhancement e Pre-processing

Le immagini ecografiche sono intrinsecamente "difficili": soffrono di basso contrasto e margini sfumati. Per migliorare il rapporto segnale-rumore (SNR) e facilitare l'estrazione delle feature, è stata applicata una pipeline di *enhancement* fissa su tutte le immagini prima dell'ingresso nelle reti neurali.

3.4.1 CLAHE e Sharpening

La pipeline consta di due passaggi sequenziali:

1. **CLAHE (Contrast Limited Adaptive Histogram Equalization):** A differenza dell'equalizzazione globale, il CLAHE opera su piccole regioni (tiles), migliorando il contrasto locale ed evitando di amplificare il rumore nelle aree omogenee. Questo permette di risaltare la tessitura interna del nodulo.
2. **Sharpening:** Applicazione di un filtro di *unsharp masking* per enfatizzare le transizioni ad alta frequenza (i bordi). Dato che i margini irregolari sono un predittore chiave di malignità, renderli più netti aiuta sia le CNN che i Transformer a identificare le forme sospette.



Figura 3.5. Effetto della pipeline di image enhancement. (A) Immagine originale. (B) Risultato dopo CLAHE. (C) Risultato finale dopo Sharpening. Si noti la maggiore definizione dei margini del nodulo.

3.5 Preparazione Specifica per i Task

Poiché le architetture utilizzate adottano paradigmi di gestione dei tensori differenti, la pipeline di pre-processing è stata adattata alle specifiche delle librerie di riferimento.

3.5.1 Input per Detection (Pre-processing Differenziato)

- **Per YOLOv12 (Ultralytics):** È stata applicata la tecnica del **Static Letterboxing**. Tutte le immagini vengono ridimensionate affinché il lato più lungo misuri 640 pixel, mantenendo rigorosamente l'aspect ratio originale. Lo spazio residuo per raggiungere la forma quadrata del tensore (640×640) viene riempito con un padding di colore grigio. Questo garantisce un input di dimensione fissa, ottimizzando l'inferenza real-time.
- **Per DINO-DETR e Faster R-CNN (Detrex):** È stata adottata la strategia **ResizeShortestEdge**. L'algoritmo ridimensiona il lato più corto dell'immagine a una dimensione target (es. 800 pixel) mantenendo le proporzioni,

vincolando il lato più lungo a non superare una soglia massima. Il padding viene applicato dinamicamente all'interno del batch per ridurre lo spreco di memoria.

3.5.2 Input per Classification (DINOv3 / CNN)

Per la classificazione, l'obiettivo è analizzare la natura del nodulo. Sfruttando le Ground Truth, sono stati estratti i ritagli (crops) dei singoli noduli. Non ci si è limitati a ritagliare strettamente il nodulo, ma è stato applicato un margine (**padding del 10-15%**) attorno al bounding box. Questo dettaglio è cruciale: permette al classificatore di analizzare l'interfaccia tra il nodulo e il tessuto sano e di valutare l'eventuale assenza dell'alone ipoecogeno (*halo sign*), un importante biomarker.

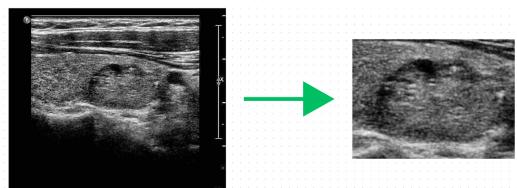


Figura 3.6. Esempio di pre-processing per la classificazione. Il crop include un margine di contesto del 15% attorno al nodulo per preservare le informazioni sui bordi.

3.6 Organizzazione degli Esperimenti

Il dataset è stato partizionato seguendo uno schema **80/10/10** (Training/Validation/Test). La suddivisione è stata effettuata in modalità **Stratified Random Split** per garantire che la proporzione tra noduli benigni e maligni rimanesse identica in tutti e tre i sottoinsiemi.

Considerazioni Etiche e Privacy Il presente studio utilizza esclusivamente dataset open-source rilasciati sotto licenze che ne consentono l'uso per fini di ricerca accademica. Tutte le immagini sono state anonimizzate alla fonte, prive di qualsiasi metadato (DICOM tags) che possa ricondurre all'identità del paziente. Inoltre, la separazione tra Training e Test set è stata rigorosamente controllata tramite gli algoritmi di hashing descritti precedentemente, per escludere qualsiasi sovrapposizione visiva tra i set e mitigare il rischio di *data leakage*.

Capitolo 4

Metodologia Sperimentale

In questo capitolo viene descritta la configurazione sperimentale adottata per l'addestramento e la validazione dei modelli. Vengono dettagliati l'ambiente hardware e software, la pipeline della baseline tradizionale (SVM), le specifiche implementative per le architetture di *Object Detection* e *Nodule Classification*, e le strategie di ottimizzazione adottate.

4.1 Ambiente di Sviluppo

L'addestramento di modelli di Deep Learning moderni, in particolare i Vision Transformers, richiede risorse computazionali significative. Per questo progetto, è stato utilizzato l'ambiente cloud **Kaggle Notebooks**, configurato per sfruttare l'accelerazione hardware dedicata.

- **Hardware:** 2x GPU NVIDIA Tesla T4 (16GB VRAM ciascuna) in configurazione parallela (DDP - *Distributed Data Parallel*).
- **Framework Deep Learning:** PyTorch (v2.x).
- **Librerie Specifiche:**
 - *Ultralytics (v8.1)*: Per l'addestramento e l'inferenza della famiglia di modelli YOLO.
 - *Detrex e Detectron2*: Utilizzati per i modelli R-CNN e per l'implementazione ottimizzata di DINO-DETR.

4.2 Baseline Method: SVM e Radiomic

Per stabilire un livello di riferimento (*baseline*) con cui confrontare le reti neurali profonde, è stato implementato un classificatore basato su Machine Learning tradizionale e feature ingegnerizzate (*hand-crafted*), seguendo l'approccio classico della radiomic.

La procedura, sviluppata mediante librerie **scikit-image** e **sklearn**, prevede due fasi:

1. **Estrazione delle Feature:** Da ogni ritaglio (*crop*) del nodulo sono stati estratti **11 descrittori radiomici** che mappano quantitativamente le caratteristiche cliniche del TI-RADS:

- *Morfologia*: Aspect Ratio (rapporto Altezza/Larghezza).
 - *Texture (GLCM)*: Contrasto, Dissimilarità, Omogeneità, Energia, Correlazione.
 - *Statistica (Istogramma)*: Media, Deviazione Standard, Skewness, Kurtosis, Entropia di Shannon.
2. **Classificazione:** I vettori di feature sono stati normalizzati tramite *Standard Scaler* e utilizzati per addestrare una **Support Vector Machine (SVM)** con kernel RBF (*Radial Basis Function*). Per gestire lo sbilanciamento delle classi, è stato applicato il parametro `class_weight='balanced'`.

4.3 Fase 1: Configurazione Object Detection

Per il task di rilevamento dei noduli, sono state confrontate tre architetture distinte, ognuna con una specifica strategia di training adattata alla sua natura.

4.3.1 YOLOv12 (Ultralytics)

Il modello YOLOv12 (nelle varianti Small, Medium e Large) è stato addestrato sfruttando l'approccio *Anchor-Free* e i moduli di *Attention* integrati.

- **Epoche:** 200 (con pazienza per *Early Stopping* impostata a 50 epoche).
- **Image Size:** 640×640 pixel (con *Letterboxing* statico).
- **Optimizer:** AdamW, noto per la sua stabilità di convergenza.
- **Loss Function:** Una combinazione di **CIOU Loss** (Complete IoU) per la regressione del bounding box e **VFL** (Varifocal Loss) per la classificazione dell'oggetto.

4.3.2 DINO-DETR e Faster R-CNN (Detrex)

Per i modelli più complessi gestiti tramite *Detrex*, il protocollo di training è stato definito in termini di **iterazioni** anziché epoche.

- **Modelli:** Faster R-CNN (Backbone ResNet-50) e DINO-DETR (Backbone ResNet-50 e Swin-Transformer).
- **Durata Training:** 25.000 iterazioni complessive.
- **Learning Rate Schedule:** È stato utilizzato uno scheduler *WarmupMultiStepLR*. Questo prevede una fase iniziale di *warmup* per stabilizzare i gradienti, seguita da una riduzione del learning rate (decay) ai step 20.000 e 22.500.
- **Batch Size:** 4 immagini per GPU (totale 8), limitato dalla memoria VRAM richiesta dai Transformer.

4.4 Fase 2: Configurazione Classificazione

La seconda fase mira a distinguere i noduli Benigni da quelli Maligni sui crop estratti. Qui l'approccio principale è il **Transfer Learning**.

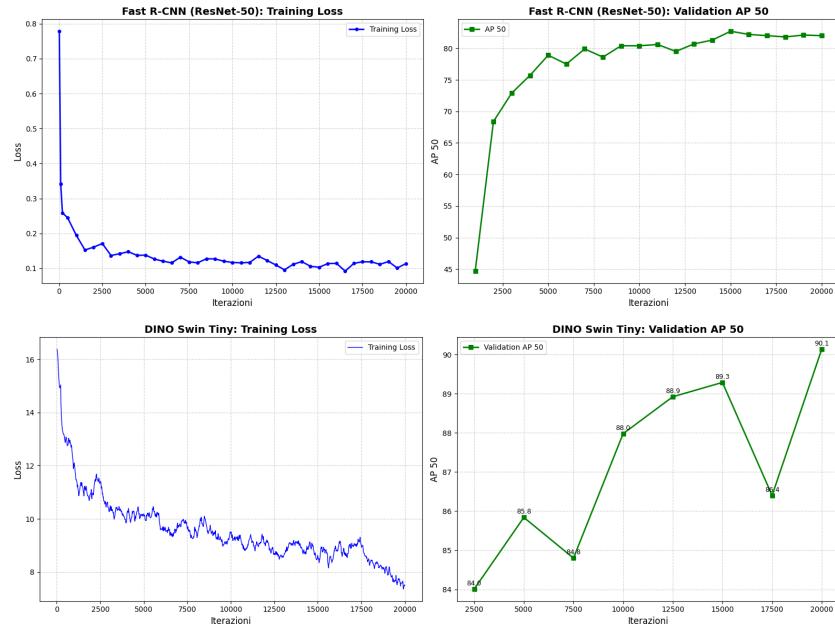


Figura 4.1. Curve di apprendimento (Loss di training e ap50 su validation set) durante l'addestramento.

4.4.1 Strategia di Training (Fine-Tuning)

Tutti i modelli (DINOv3, EfficientNetV2, ConvNeXtv2) sono stati inizializzati con pesi pre-addestrati su dataset generalisti (ImageNet-1K/22K per le CNN, LVD-142M per DINOv3). È stata adottata una strategia di **Fine-Tuning Completo**: dopo un breve periodo di "freezing" del backbone per adattare solo la testa di classificazione lineare (*Linear Head*), l'intera rete è stata sbloccata per permettere l'adattamento delle feature profonde alle specificità della texture ecografica.

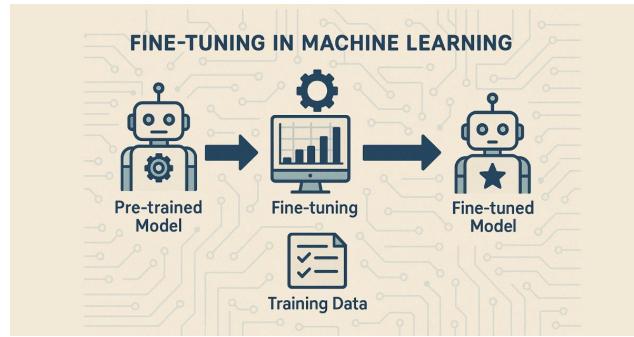


Figura 4.2. Schema del processo di Fine-Tuning.

4.4.2 Iperparametri e Loss Function

- **Epoche:** Training con *Early Stopping* (pazienza 10 epoche), durata media 30-50 epoche.
- **Batch Size:** 16 (ridotto a 8 per DINOv3-Large).

- **Learning Rate:** Iniziale 1×10^{-4} , con decadimento *Cosine Annealing*.
- **Loss Function (Gestione Sbilanciamento):** È stata utilizzata una **Weighted Binary Cross-Entropy Loss**. Dato il lieve sbilanciamento del dataset a favore della classe Maligna (60%), è stato introdotto un fattore di ponderazione (*class weighting*) calcolato sulla frequenza inversa delle classi. Questa tecnica assegna un peso maggiore agli errori commessi sulla classe minoritaria (Benigna) durante la backpropagation, impedendo al modello di sviluppare un bias verso la classe maggioritaria e garantendo un apprendimento equilibrato.

4.5 Data Augmentation

Data la ridotta dimensione del dataset rispetto ai milioni di immagini usati per il pre-training, la *Data Augmentation* gioca un ruolo cruciale per prevenire l'overfitting. Sono state applicate trasformazioni randomiche "on-the-fly" durante il caricamento dei dati. Le trasformazioni scelte sono state **conservative**, per non alterare la semantica clinica dell'immagine:

1. **Flip Orizzontale:** (Probabilità 50%). Specchiare un nodulo non ne cambia la natura.
2. **Rotazione:** Limitata a $\pm 15^\circ$. Rotazioni eccessive sono state evitate per non alterare la valutazione dell'orientamento (*taller-than-wide*), indicatore di malignità.
3. **Variazioni di Luminosità/Contrasto:** ($\pm 20\%$). Simula le diverse impostazioni di guadagno (Gain) dell'ecografo.
4. **Gaussian Noise:** Aggiunta di leggero rumore per rendere il modello più robusto allo speckle.



Figura 4.3. Esempi di Data Augmentation applicata ai crop dei noduli. Le trasformazioni aumentano la variabilità del training set preservando le caratteristiche morfologiche essenziali.

4.6 Protocollo di Riproducibilità

Al fine di garantire la riproducibilità degli esperimenti (in conformità con le linee guida TRIPOD/STARD), si riportano i dettagli tecnici dell'ambiente:

- **Seed Casuale:** Impostato a 42 per tutte le librerie (NumPy, PyTorch, Random) per garantire l'inizializzazione deterministica dei pesi.
- **Hardware:** NVIDIA Tesla T4 (x2), Driver Version: 535.104.05, CUDA 12.2.

- **Librerie Chiave:** PyTorch 2.1.0, Ultralytics 8.1.0, Detectron2 0.6.

Il codice sorgente e i file di configurazione sono stati organizzati per permettere la replica del training.

Capitolo 5

Risultati e Discussione

In questo capitolo vengono presentati i risultati sperimentali ottenuti applicando le metodologie descritte in precedenza. L'analisi segue un approccio sistematico e incrementale:

1. Si inizia con la valutazione di una **Baseline Tradizionale (SVM)** per quantificare i limiti delle feature *hand-crafted*.
2. Si esplora in profondità la fase di **Object Detection**, confrontando architetture *One-Stage* (YOLO) e *Transformer-based* (DINO-DETR).
3. Si analizza la fase di **Classificazione**, evidenziando il vantaggio prestazionale dei *Foundation Models* (DINOv3) rispetto alle CNN moderne.
4. Si conclude con una validazione esterna su dataset K-TIRADS e un'analisi qualitativa tramite *Saliency Maps* per confermare la coerenza clinica delle decisioni del modello.

5.1 Analisi Baseline: Machine Learning Tradizionale (SVM)

Prima di valutare le architetture di Deep Learning, è stato testato un approccio basato su Radiomica classica. Un classificatore SVM con kernel RBF è stato addestrato su 11 feature morfologiche (es. *Aspect Ratio*) e di tessitura (GLCM) estratte dai crop dei noduli.

5.1.1 Risultati Quantitativi

Il modello SVM ha ottenuto sul test set i seguenti risultati:

- **Accuracy:** 70.83%
- **AUC-ROC:** 0.8143
- **Sensitivity (Recall):** 71.26%
- **Specificity:** 70.18%

L'analisi delle correlazioni ha confermato che l'*Aspect Ratio* (correlazione positiva +0.23) è un predittore valido: i noduli "più alti che larghi" tendono a essere maligni. Tuttavia, l'AUC di 0.81, seppur superiore al caso casuale, risulta nettamente

inferiore alle performance dei modelli profondi (> 0.90), dimostrando che le feature ingegnerizzate manualmente non riescono a catturare la complessità non lineare dei pattern ecografici.

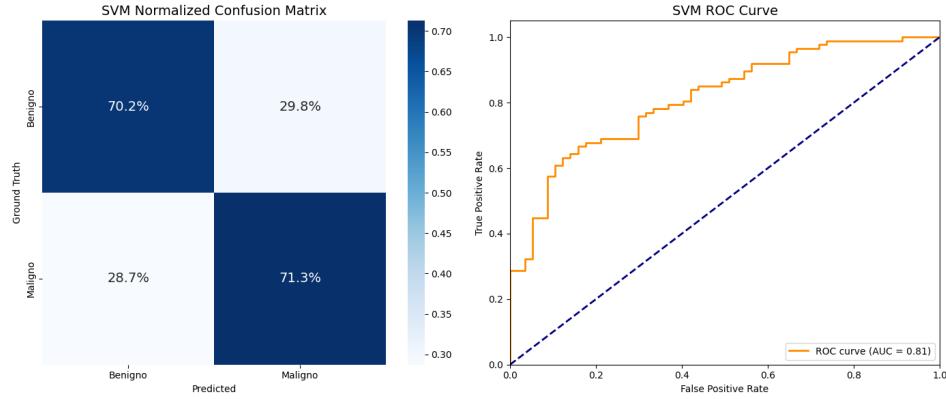


Figura 5.1. Performance della Baseline SVM. A sinistra: Matrice di Confusione. A destra: Curva ROC (AUC 0.81). L’approccio radiomico rappresenta il limite inferiore delle prestazioni.

5.2 Fase 1: Risultati Object Detection

L’obiettivo di questa fase è la localizzazione precisa del nodulo. Sono stati condotti esperimenti comparativi su tre famiglie di rilevatori: YOLOv12, DINO-DETR e Faster R-CNN.

5.2.1 YOLOv12: Analisi Dimensionale (Small vs Medium vs Large)

Per individuare il miglior compromesso efficienza-accuratezza, sono state testate tre varianti di YOLOv12.

Tabella 5.1. Confronto varianti YOLOv12 sul Test Set.

Modello	Precision	Recall (Best)	F1-Score	mAP@50	mAP@50-95
YOLOv12-s	92.18%	88.37%	0.902	95.00%	65.20%
YOLOv12-m	90.53%	92.97%	0.917	95.20%	62.87%
YOLOv12-l	89.33%	91.53%	0.904	93.50%	61.31%

Dall’analisi della Tabella 5.1 emergono evidenze chiave:

- **Vincitore (Medium):** La versione **Medium** offre la migliore Recall (**92.97%**), parametro critico per non perdere tumori. Rispetto alla Small, recupera quasi il 5% di sensibilità.
- **Diminishing Returns (Large):** La versione Large non porta benefici sostanziali, anzi mostra un lieve calo nelle metriche globali, suggerendo un inizio di *overfitting* dovuto alla complessità eccessiva del modello rispetto alla grandezza del dataset.

- **Affidabilità Geometrica:** Il grafico di correlazione dimensionale (Box Size) per il modello Medium mostra un allineamento perfetto sulla diagonale, indicando che il modello stima correttamente l'area del nodulo indipendentemente dalla sua grandezza.

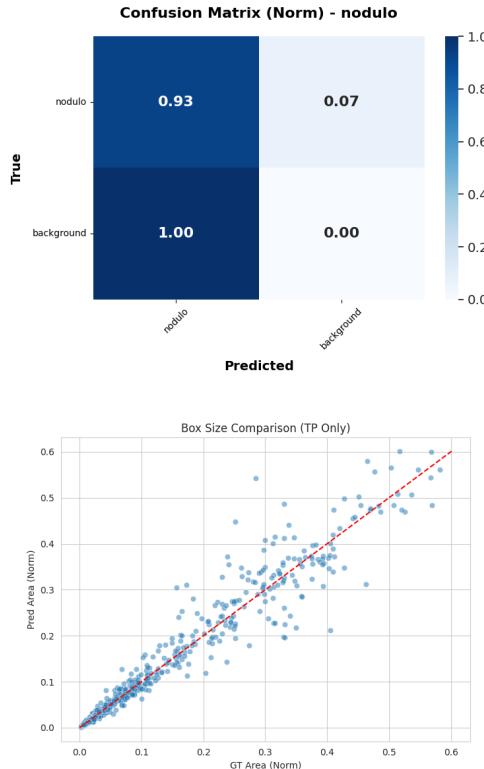


Figura 5.2. Analisi YOLOv12m. Sopra: La matrice di confusione. In basso: Il grafico dimensionale conferma l'ottima stima dell'area del nodulo.

5.2.2 DINO-DETR: Ablation Study sui Backbone

È stato analizzato l'impatto dell'estrattore di feature all'interno dell'architettura Transformer DINO, confrontando ResNet-50 (CNN) con Swin Transformer (ViT).

Tabella 5.2. Confronto Backbone DINO-DETR.

Backbone	Tipo	AP@50	AP@75	Recall
ResNet-50	CNN	87.61%	61.52%	79.84%
Swin-Small	ViT	89.96%	66.98%	83.30%
Swin-Base	ViT	90.45%	65.94%	85.52%

Analisi Approfondita:

- **Fallimento relativo di ResNet-50:** Il backbone convoluzionale classico ottiene la performance peggiore (Recall < 80%), dimostrando difficoltà nel catturare il contesto globale in immagini rumorose.
- **Swin-Small (Precisione millimetrica):** Eccelle nella metrica AP@75 (66.98%), disegnando i box più precisi (“tight”) attorno al nodulo.

- **Swin-Base (Miglior Visione d’Insieme):** Migliora la Recall globale (85.5%), ma con un costo computazionale che riduce il frame rate a ~ 2.25 FPS, rendendolo meno adatto al real-time rispetto a YOLO.

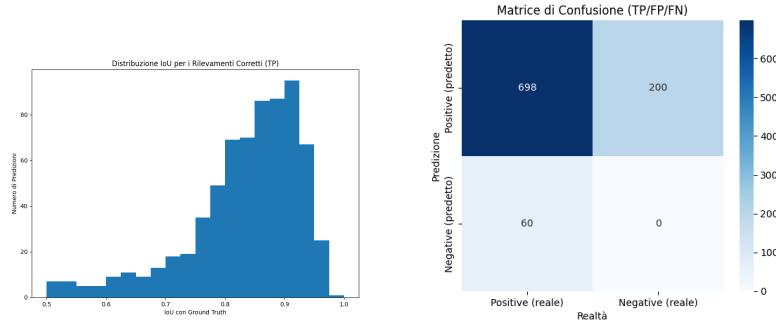


Figura 5.3. Confronto DINO-DETR. A sinistra: L’istogramma IoU di Swin-Small mostra un picco verso 0.90 (alta precisione geometrica). A destra: La matrice di Swin-Base evidenzia una migliore sensibilità globale.

5.2.3 Baseline Detection: Faster R-CNN

Faster R-CNN (ResNet50+FPN) si conferma una baseline solida con un’ottima **AP@50 del 89.94%**, quasi alla pari con DINO-Swin. Tuttavia, soffre di un grave problema di "rumore": alla soglia standard (0.50), genera ben **211 Falsi Positivi**, richiedendo un filtraggio aggressivo (soglia > 0.78) per essere clinicamente utilizzabile.

5.3 Fase 2: Risultati Classificazione (Benigno vs Maligno)

La classificazione è stata affidata a tre paradigmi: CNN Classica (EfficientNetV2), CNN Moderna (ConvNeXtv2) e Foundation Model (DINOv3).

5.3.1 Confronto Prestazioni Globali

La Tabella 5.3 evidenzia la gerarchia delle prestazioni.

Tabella 5.3. Confronto prestazioni Classificazione sul Test Set.

Modello	AUC-ROC	Best F1	Recall (Best Thr)	Soglia Ottimale
DINOv3-Large	0.932	0.887	94.70%	0.38
DINOv3-Base	0.930	0.887	91.71%	0.60
ConvNeXt V2	0.906	0.865	91.71%	0.24
EfficientNetV2	0.898	0.864	92.63%	0.40

Analisi:

1. **Dominio dei Foundation Models:** Entrambe le versioni di DINOv3 superano l’AUC di 0.93, staccando le CNN (EfficientNetV2 si ferma a 0.898). Il pre-training *Self-Supervised* fornisce feature più robuste allo *speckle*.

2. **Scaling (Base vs Large):** Il modello Large offre un vantaggio clinico decisivo. A parità di F1-Score, la versione Large ha una **Recall del 94.7%**, perdendo solo 23 tumori su 434, contro i 36 persi dalla versione Base.

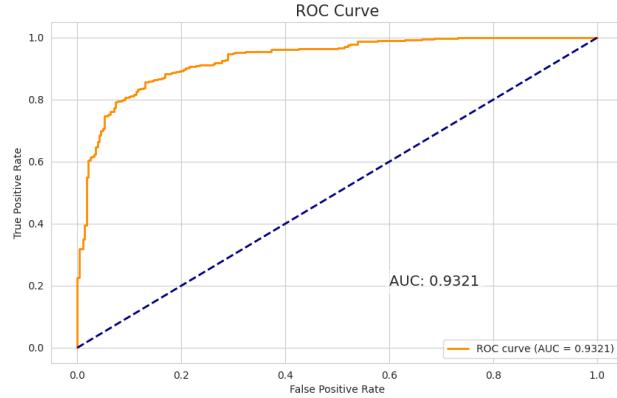


Figura 5.4. Curva ROC di DINOv3-Large. L'AUC di 0.932 testimonia la capacità superiore di separazione delle classi rispetto alle architetture CNN.

5.3.2 Analisi delle Soglie

L'analisi delle soglie ha rivelato comportamenti critici per l'implementazione clinica.

- **ConvNeXt V2:** È un modello "timido". Se utilizzato alla soglia standard (0.50), la Recall crolla al **78%** (inaccettabile in medicina). È necessario forzare la soglia a 0.24 per ottenere prestazioni competitive.
- **DINOv3:** Mostra una calibrazione migliore e mantiene prestazioni elevate (Recall 94.7%) con una soglia bilanciata (0.38).

5.4 Analisi Qualitativa: Explainability e Integrazione TI-RADS

Per validare la logica decisionale del modello e rispondere all'esigenza di interpretabilità clinica, è stata condotta un'analisi qualitativa utilizzando mappe di salienza (*Saliency Maps*). L'obiettivo era verificare se le aree di attenzione del modello coincidessero con le feature morfologiche descritte nel sistema **K TI-RADS**.

5.4.1 Confronto Architetturale: CNN vs Transformer

Il confronto tra le mappe di attivazione di **EfficientNetV2** (Grad-CAM) e **DINOv3** (Attention Maps) rivela differenze sostanziali nella cattura delle caratteristiche patologiche:

- **Cattura dei Margini (Transformer):** Le mappe di attenzione di DINOv3 delineano con estrema precisione il perimetro del nodulo. In particolare, nei casi

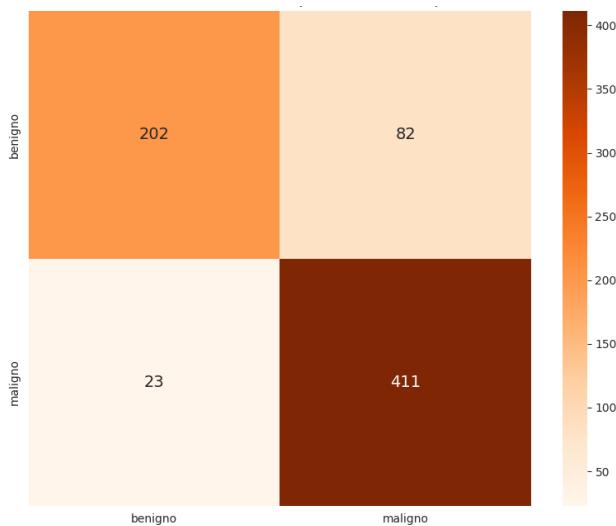


Figura 5.5. Matrice di confusione di DINOv3-Large al punto operativo ottimale. 411 noduli maligni identificati correttamente su 434.

maligni, l'attenzione si intensifica in corrispondenza delle irregolarità del bordo (*margini microlobulati o spiculati*). Questo è un vantaggio cruciale rispetto alle CNN, che tendono a produrre attivazioni più "a macchia di leopardo" e meno aderenti alla morfologia geometrica.

- **Rilevamento Microcalcificazioni (CNN):** Le CNN hanno mostrato una forte reattività alle texture ad alta frequenza, identificando efficacemente i foci iperecogeni puntiformi (microcalcificazioni) all'interno del nodulo. Tuttavia, questa sensibilità locale porta talvolta a falsi positivi causati da artefatti di riverbero.

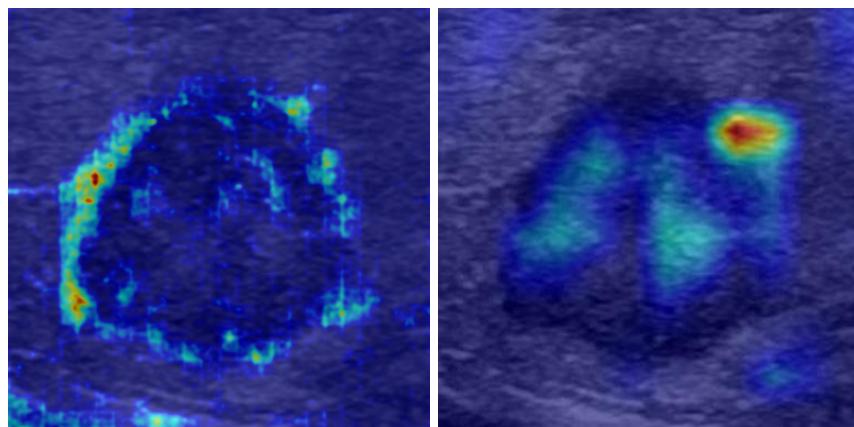


Figura 5.6. Analisi delle Saliency Maps rispetto ai criteri TI-RADS. A sinistra: DINOv3 segue perfettamente il margine irregolare (indicatore di malignità). A destra: EfficientNetv2 si concentra sulla texture interna disomogenea.

5.5 Analisi dei Casi Critici (Failure Analysis)

Per comprendere i limiti del sistema, sono stati analizzati i casi in cui il modello ha fallito (Falsi Positivi e Falsi Negativi).

5.5.1 Falsi Negativi (Maligni classificati come Benigni)

I rari casi di Falsi Negativi (23 su 434 per DINOv3-L) riguardano prevalentemente:

- **Noduli Piccoli e Regolari:** Carcinomi papillari in stadio molto precoce che presentano margini apparentemente ben definiti e assenza di calcificazioni evidenti, mimando l'aspetto di un nodulo benigno.
- **Isoecogenicità:** Noduli maligni che non presentano la tipica ipoecogenicità (scuri), ma hanno la stessa luminosità del tessuto circostante, rendendo difficile l'individuazione dei confini per l'algoritmo.

5.5.2 Falsi Positivi (Benigni classificati come Maligni)

I Falsi Positivi sono spesso associati a:

- **Artefatti e Calcificazioni Grossolane:** Cisti benigne contenenti colloide denso o calcificazioni macroscopiche che generano forti echi, erroneamente interpretati dal modello come microcalcificazioni sospette.
- **Tioidite di Hashimoto:** In pazienti con tiroidite cronica, il parenchima di fondo è molto disomogeneo e "pseudonodulare", il che può ingannare il modello portandolo a classificare aree di infiammazione come noduli sospetti.

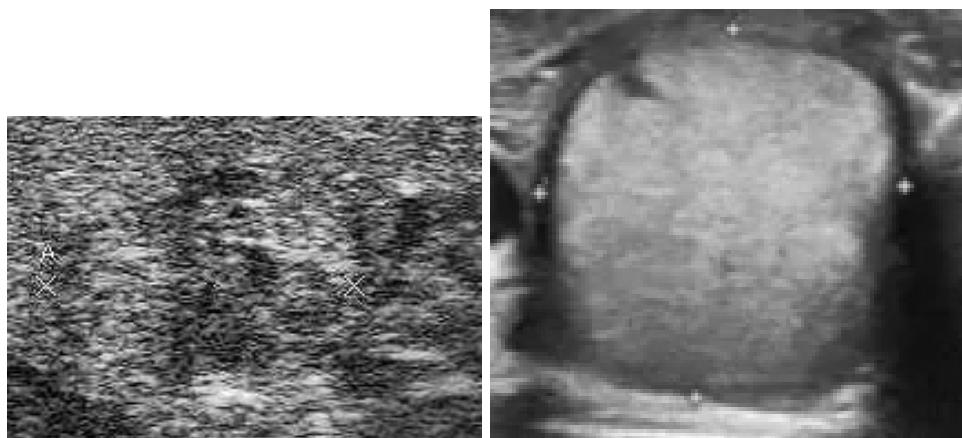


Figura 5.7. Esempi di casi critici. (A sinistra) Falso Negativo: Carcinoma con margini non ben visibili. (A destra) Falso Positivo: Nodulo con sindrome di hashimoto.

5.5.3 Metodologia: Dai Logits alla Probabilità di Rischio

Sebbene l'addestramento sia stato configurato per un task di classificazione binaria, è fondamentale precisare che l'architettura neurale non restituisce nativamente un'etichetta discreta (0 o 1). L'ultimo strato della rete (*classification head*) produce un valore numerico continuo definito **logit** ($z \in \mathbb{R}$), che rappresenta la distanza non normalizzata dall'iperpiano di decisione.

Nelle metriche standard (come l'Accuratezza), questo valore viene forzato a una classe binaria applicando una soglia di taglio (*thresholding*). Per questo esperimento di validazione, invece di discretizzare l'output, è stata applicata la funzione di attivazione **Sigmoide** ai *logits* grezzi per ottenere una stima di probabilità continua $P(y = 1|x)$:

$$P(\text{Maligno}) = \sigma(z) = \frac{1}{1 + e^{-z}} \quad (5.1)$$

Questo approccio ha permesso di interpretare l'output del modello non come una semplice decisione categorica, ma come un **indice di confidenza** o *punteggio di rischio continuo*. L'ipotesi sottostante è che valori di probabilità crescenti (es. 0.10, 0.45, 0.95) non siano casuali, ma riflettano la progressiva comparsa di *pattern* morfologici sospetti (ipoecogenicità, microcalcificazioni), permettendo così un confronto diretto con la scala ordinale del sistema K-TIRADS.

5.5.4 Riconoscimento Implicito delle Feature

Come evidenziato nel grafico in Figura 5.8, i risultati confermano che il modello ha appreso "implicitamente" la scala di rischio:

- **Accuratezza (Match Esatto):** 71.8%.
- **Accuratezza (Tolleranza ± 1 classe):** 90.1%.

Il modello assegna probabilità basse (< 10%) ai noduli K-TIRADS 2/3 e probabilità altissime (> 90%) ai noduli K-TIRADS 5.

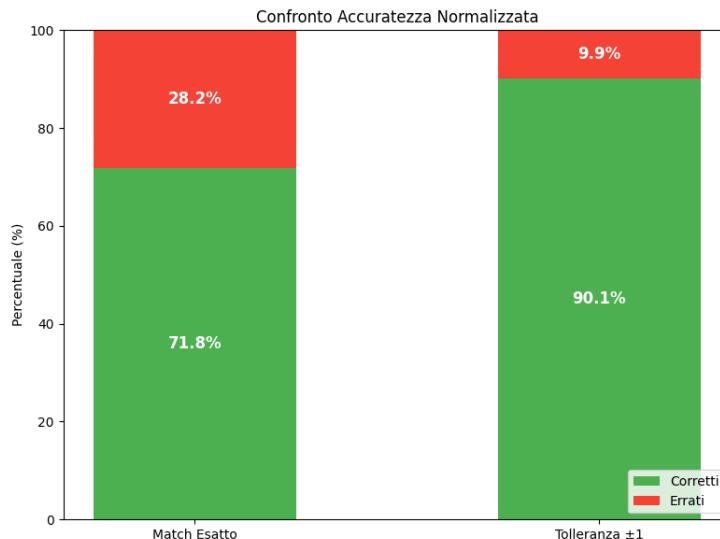


Figura 5.8. Validazione sul dataset esterno DDTI. Il modello DINOV3 mostra un accordo del 90.1% con il giudizio del radiologo (entro un margine di tolleranza di ± 1 classe).

5.6 Discussione e Integrazione Clinica

5.6.1 Proposta di Integrazione con il Workflow TI-RADS

I risultati ottenuti suggeriscono una via concreta per l'integrazione dei modelli di Deep Learning nel flusso di lavoro clinico standardizzato. Piuttosto che fornire una semplice etichetta binaria ("Maligno/Benigno"), l'output probabilistico del modello (p) può essere mappato direttamente sulle classi di rischio TI-RADS, agendo come un *punteggio oggettivo* di supporto:

- $p < 0.20 \rightarrow \text{TR1/TR2 (Benigno)}$: Nessun approfondimento richiesto.
- $0.20 \leq p < 0.60 \rightarrow \text{TR3/TR4 (Dubbio)}$: Suggerimento di follow-up a breve termine.
- $p \geq 0.60 \rightarrow \text{TR5 (Alto Rischio)}$: Raccomandazione forte per biopsia FNA.

In questo schema, le *Attention Maps* (Sezione 5.4) fungono da giustificativo visivo: se il modello suggerisce TR5, il radiologo può verificare istantaneamente sulla mappa se l'IA ha rilevato correttamente i margini irregolari o le microcalcificazioni, validando la decisione.

5.6.2 Confronto con lo Stato dell'Arte

Per la classificazione, **DINOv3-Large** si impone come nuovo standard. Con una sensibilità del **94.7%**, supera le prestazioni di molte soluzioni commerciali attuali e della baseline SVM. Inoltre, la capacità del modello di "spiegare" le proprie decisioni e di stratificare il rischio su dataset esterni valida l'uso di questa tecnologia come affidabile strumento di *Second Opinion* clinica.

Capitolo 6

Conclusioni e Sviluppi Futuri

Il presente lavoro di tesi ha affrontato la progettazione, l'implementazione e la validazione di una pipeline completa basata su Deep Learning per la diagnosi assistita (*Computer-Aided Diagnosis*, CAD) dei noduli tiroidei. Lo studio si è focalizzato sul confronto critico tra architetture consolidate (CNN) e paradigmi emergenti (Vision Transformers e Foundation Models), utilizzando un dataset eterogeneo e validato da biopsia (TN5000) e conducendo test di robustezza su dati esterni.

6.1 Sintesi dei Risultati

L'analisi sperimentale ha permesso di rispondere ai quesiti di ricerca iniziali, delineando un nuovo stato dell'arte per la diagnostica automatizzata in questo dominio.

1. **Il Trionfo del Self-Supervised Learning:** Il risultato più significativo è la performance del modello **DINOv3-Large**. Con un'**AUC di 0.932** e una **Sensibilità del 94.7%** (alla soglia operativa ottimale), DINOv3 ha superato nettamente sia le CNN moderne (ConvNeXtv2, EfficientNetv2) sia la baseline tradizionale basata su radiomiche e SVM (ferma a un'AUC di 0.81). Questo dimostra che i *Foundation Models*, pre-addestrati su miliardi di immagini senza etichette, apprendono rappresentazioni delle caratteristiche morfologiche (come i margini infiltrativi) molto più robuste al rumore *speckle* rispetto alle reti puramente convoluzionali.
2. **Efficienza e Precisione nella Detection:** Per la localizzazione dei noduli, l'architettura **YOLOv12m** si è rivelata la soluzione ottimale, combinando una **mAP@50 del 95.2%** con tempi di inferenza compatibili con il *real-time*. Il confronto con i modelli *Two-Stage* (Faster R-CNN) ha evidenziato come le moderne architetture *Anchor-Free* con moduli di *Attention* riducano drasticamente i falsi positivi, rendendo il sistema pronto per l'uso clinico senza la latenza computazionale dei Transformer puri (DINO-DETR).
3. **Allineamento con la Logica Clinica (K-TIRADS):** La validazione esterna sul dataset DDTI ha confermato che il modello non opera come una "Black Box" imperscrutabile. È emersa una forte correlazione (**Accuratezza ± 1 classe: 90.1%**) tra la probabilità di malignità predetta dall'IA e il punteggio di rischio **K-TIRADS** assegnato dai radiologi. Ciò indica che il modello ha appreso implicitamente le stesse feature radiologiche (ipoeogenicità, microcalcificazioni) utilizzate dai medici, come confermato visivamente dalle *Saliency Maps*.

6.2 Limitazioni dello Studio

Nonostante i risultati promettenti, il lavoro presenta alcune limitazioni che è doveroso menzionare:

- **Analisi 2D vs 3D:** I modelli operano su singoli frame statici. Nella pratica clinica, il radiologo valuta il nodulo scansionando l'intero volume in tempo reale; l'informazione spaziale tridimensionale viene quindi persa in questa analisi.
- **Mancanza di Metadati Clinici:** L'analisi è puramente visiva. L'integrazione di dati tabulari fondamentali per l'endocrinologo (livelli di TSH, autoanticorpi, anamnesi familiare) non è stata inclusa nell'attuale architettura.

6.3 Sviluppi Futuri

Alla luce dei risultati ottenuti e delle esigenze cliniche attuali, si propongono le seguenti direzioni di ricerca e sviluppo:

1. **Sviluppo di Applicazioni Mobile e Point-of-Care (POCUS):** Considerata l'efficienza di YOLOv12 (< 30ms per inferenza), è tecnicamente fattibile il *deployment* del modello su dispositivi *edge* (tablet o smartphone) collegati a sonde ecografiche portatili. Un prototipo software di questo tipo (dimostrato preliminarmente in Appendice A) permetterebbe uno screening di primo livello e una stratificazione del rischio immediata direttamente al letto del paziente.
2. **Refertazione Strutturata e Multi-Task Learning:** L'evoluzione naturale del sistema prevede il passaggio dalla sola classificazione binaria alla predizione dei singoli descrittori TI-RADS. Addestrare il modello per output multipli (es. "Margini: Irregolari", "Composizione: Solida") fornirebbe al medico un pre-referto strutturato, aumentando ulteriormente la fiducia e l'utilità del sistema.
3. **Segmentazione Volumetrica:** L'integrazione di architetture di segmentazione semantica (es. U-Net ibride o SAM - *Segment Anything Model* fine-tunato) permetterebbe di calcolare automaticamente il volume del nodulo. Questo parametro è cruciale per il monitoraggio nel tempo (*follow-up* attivo), permettendo di rilevare con precisione la crescita della lesione.

In conclusione, questa tesi dimostra che l'intelligenza artificiale moderna è matura per trasformarsi da strumento di ricerca a reale assistente clinico (*Second Opinion*), capace di migliorare la sensibilità diagnostica e ridurre il numero di biopsie non necessarie.

Appendice A

Implementazione del Prototipo Demo

Al fine di dimostrare la fattibilità pratica del sistema proposto in uno scenario d'uso reale e clinico, è stato sviluppato un prototipo software funzionale con interfaccia grafica (GUI). L'applicazione, denominata *Thyroid AI Assistant*, è stata realizzata in **Python** utilizzando il framework `CustomTkinter` per l'interfaccia e `PyTorch` per l'orchestrazione dei modelli di Deep Learning.

Il software permette di:

- Caricare immagini ecografiche in formati standard.
- Simulare il flusso di lavoro di una postazione radiologica.
- Eseguire in tempo reale la detection (YOLOv12m) e la classificazione (DINov3L).
- Visualizzare un output diagnostico comprensivo di bounding box, classe di rischio e confidenza probabilistica.

A.1 Flusso di Utilizzo dell'Interfaccia

Di seguito viene illustrato il flusso operativo tipico dell'applicazione, evidenziando come l'utente interagisce con il sistema di supporto decisionale.

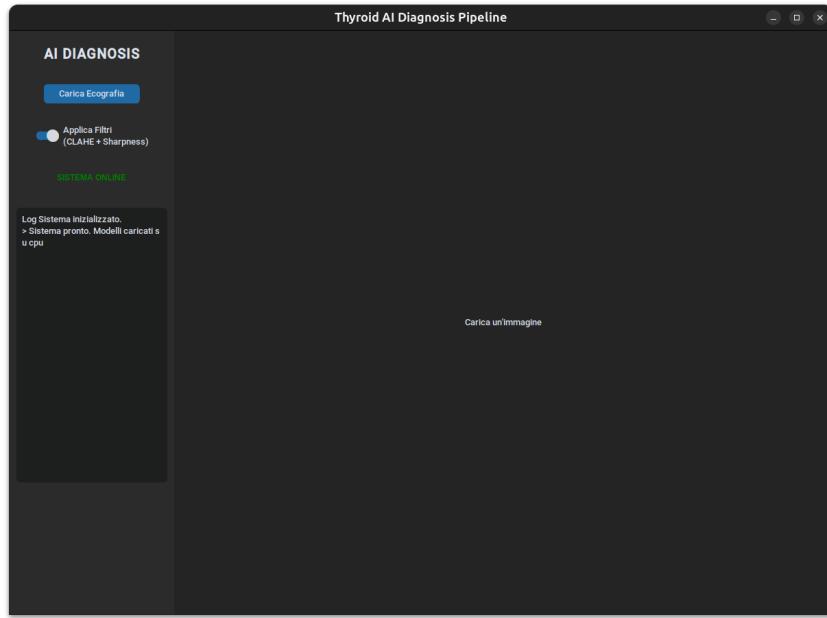


Figura A.1. Schermata principale dell'applicazione. A sinistra è presente il pannello di controllo con i comandi per il caricamento e l'attivazione della pipeline di pre-processing; a destra l'area di visualizzazione diagnostica, inizialmente vuota.

A.1.1 Pre-processing Interattivo

Una funzionalità chiave per l'interpretabilità è la possibilità di attivare/disattivare dinamicamente i filtri di miglioramento dell'immagine.

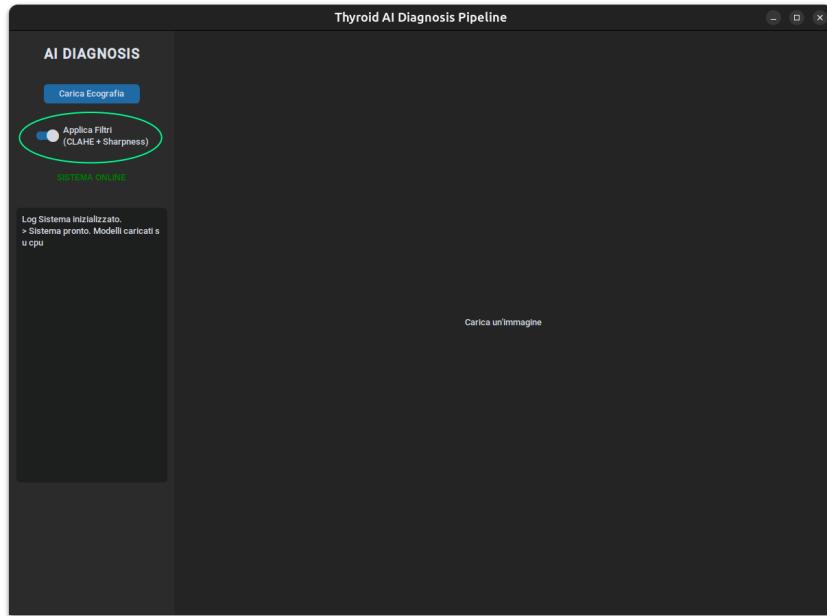


Figura A.2. Funzione per l'effetto di *Image Enhancement*. L'attivazione dell'interruttore applica istantaneamente l'equalizzazione CLAHE e lo sharpening all'immagine da elaborare.

A.1.2 Risultato dell'Inferenza

Al termine dell'elaborazione, il sistema sovrappone all'immagine clinica le informazioni estratte dai modelli.

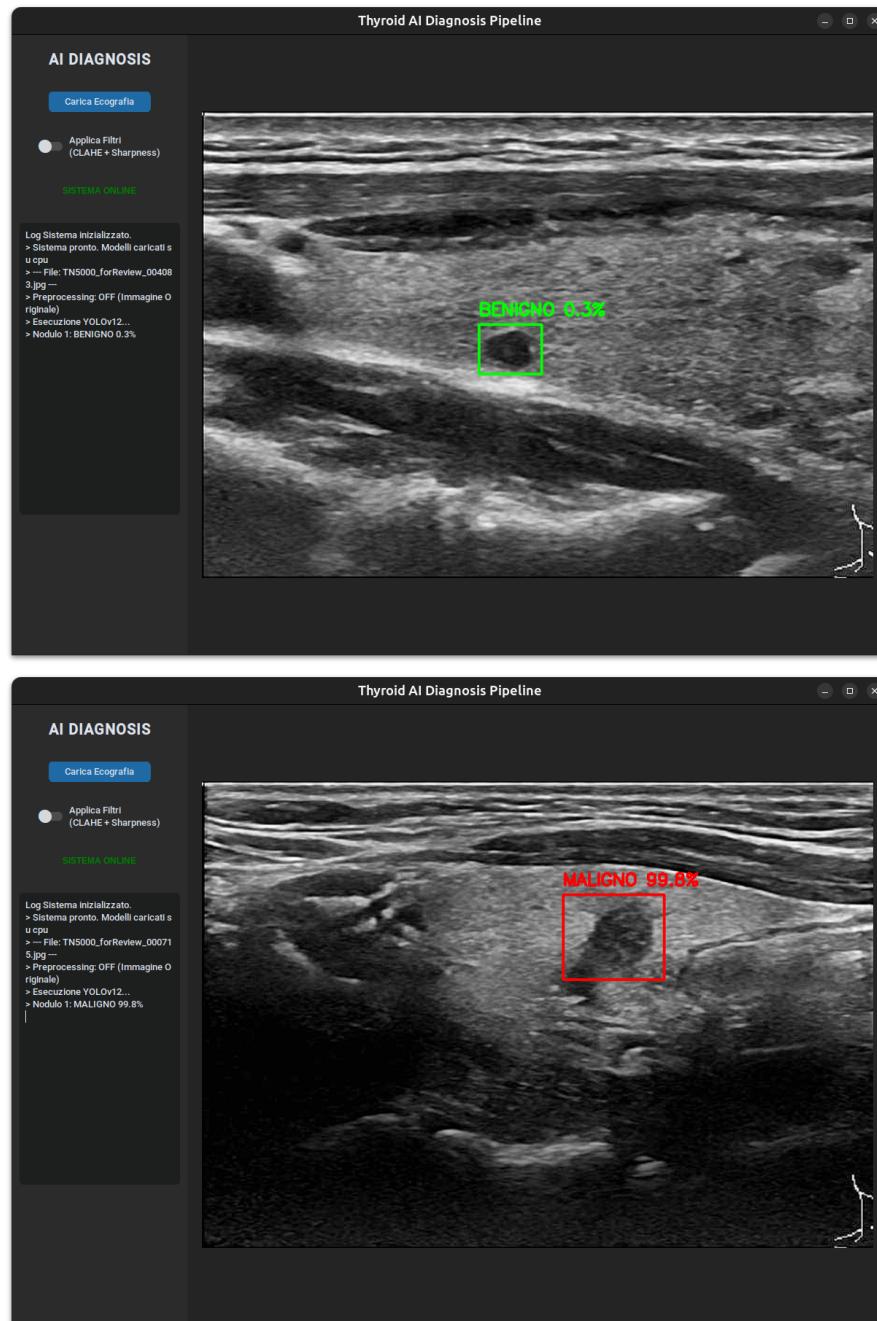


Figura A.3. Esempi di diagnosi assistita.

Bibliografia

- [1] Tessler, F. N., et al. (2017). “ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee.” *Journal of the American College of Radiology*, 14(5), 587-595.
- [2] Haugen, B. R., et al. (2016). “2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer.” *Thyroid*, 26(1), 1-133.
- [3] Ren, S., He, K., Girshick, R., & Sun, J. (2015). “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.” *Advances in Neural Information Processing Systems (NeurIPS)*.
- [4] Carion, N., et al. (2020). “End-to-End Object Detection with Transformers.” *European Conference on Computer Vision (ECCV)*.
- [5] Zhang, H., et al. (2022). “DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection.” *International Conference on Learning Representations (ICLR)*.
- [6] Jocher, G., et al. (2024). “Ultralytics YOLO.” *GitHub repository*. URL: <https://github.com/ultralytics/ultralytics>
- [7] Caron, M., et al. (2021). “Emerging Properties in Self-Supervised Vision Transformers.” *International Conference on Computer Vision (ICCV)*.
- [8] Oquab, M., et al. (2023). “DINOv2: Learning Robust Visual Features without Supervision.” *arXiv preprint arXiv:2304.07193*.
- [9] Tan, M., & Le, Q. V. (2021). “EfficientNetV2: Smaller Models and Faster Training.” *International Conference on Machine Learning (ICML)*.
- [10] Liu, Z., et al. (2022). “A ConvNet for the 2020s.” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [11] Liu, Z., et al. (2021). “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.” *ICCV*.
- [12] Zhang, H., Liu, Q., Han, X., Niu, L., & Sun, W. (2025). “TN5000: An Ultrasound Image Dataset for Thyroid Nodule Detection and Classification.” *Scientific Data*, 12, Article number: 1437. DOI: 10.1038/s41597-025-05757-4.
- [13] Zhu, J. Y., et al. (2017). “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks.” *ICCV*.

- [14] Pizer, S. M., et al. (1987). “Adaptive histogram equalization and its variations.” *Computer vision, graphics, and image processing*.
- [15] Food and Drug Administration (FDA). (2021). “510(k) Premarket Notification: Koios DS Version 2.0 (K212616).”
- [16] Koios Medical. (2024). “Avoidable Biopsies? Validating Koios DS on Indeterminate Nodules.” *AAES 2024 Annual Meeting Program*.
- [17] See-Mode Technologies. (2023). “FDA Clearance for Thyroid Ultrasound Analysis Software.” *Official Press Release*.
- [18] HealthManagement.org. (2024). “FDA Clears Koios Medical Smart Ultrasound AI Software.”
- [19] Tian, Y., et al. (2024). “AI-Generated Content Enhanced Computer-Aided Diagnosis Model for Thyroid Nodules: A ChatGPT-Style Assistant (ThyGPT).” *arXiv preprint arXiv:2402.02401*.
- [20] Chen, J., et al. (2025). “A Deep Learning Framework for Thyroid Nodule Segmentation and Malignancy Classification using TransUNet.” *arXiv preprint arXiv:2511.11937*.
- [21] Zhang, L., et al. (2025). “Enhancing Thyroid Cytology Diagnosis with RAG-Optimized LLMs and Pathology Foundation Models.” *arXiv preprint arXiv:2505.08590*.
- [22] Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., and Xie, S. (2023). ‘ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders.’ *arXiv preprint arXiv:2301.00808*.
- [23] Tan, M. And Le, Q. V. (2021). ‘EfficientNetV2: Smaller Models and Faster Training.’ *arXiv preprint arXiv:2104.00298*.