UNIVERSITÀ
POLITECNICA
DELLE MARCHE

Facoltà di Ingegneria
Corso di Laurea Magistrale in Ingegneria Informatica e dell'Automazione

---

Un approccio automatizzato basato su Vision Transformer per la valutazione ecografica del Covid-19

An automated Vision Transformer approach for Covid-19 ultrasound assessment

Candidato:
Andrian Melnic

Relatore:
Prof. Primo Zingaretti

Correlatori:
Maria Chiara Fiorentino, Ph.D
Riccardo Rosati, Ph.D

Anno Accademico 2022-2023

UNIVERSITÀ
POLITECNICA
DELLE MARCHE

Facoltà di Ingegneria
Corso di Laurea Magistrale in Ingegneria Informatica e dell'Automazione

---

Un approccio automatizzato basato su Vision Transformer per la valutazione ecografica del Covid-19

An automated Vision Transformer approach for Covid-19 ultrasound assessment

Candidato:
Andrian Melnic

Relatore:
Prof. Primo Zingaretti

Correlatori:
Maria Chiara Fiorentino, Ph.D
Riccardo Rosati, Ph.D

Anno Accademico 2022-2023

Università Politecnica delle Marche
Facoltà di Ingegneria
Corso di Laurea Magistrale in Ingegneria Informatica e dell'Automazione
Via Brecce Bianche – 60131 Ancona (AN), Italy

Ringraziamenti

Desidero esprimere la mia profonda gratitudine a tutte quelle persone che hanno reso possibile la realizzazione di questa tesi e mi hanno accompagnato in questo mio viaggio di crescita accademica e sopratutto personale. In primo luogo vorrei ringraziare la mia famiglia per avermi dato la possibilità di intraprendere questo percorso dandomi fiducia nonostante i miei alti e bassi. Grazie a mia madre Natalia per l'amore incondizionato, il sostegno, la pazienza e gli enormi sacrifici compiuti per crescermi da sola senza farmi mancare mai nulla, nonostante tutte le difficoltà che abbiamo attraversato. Un grande grazie anche a Luigi, per la sua gentilezza e affetto nel trattarmi come un figlio e sopratutto per aver aiutato mia madre supportandola e ascoltandola nei momenti di bisogno.

Un ringraziamento speciale va al Prof. Primo Zingaretti per l'opportunità di lavorare su un progetto impegnativo e stimolante e ai miei correlatori Maria Chiara e Riccardo per il loro impegno, la loro disponibilità e i loro preziosi consigli. Aver avuto l'opportunità di lavorare con voi è stato per me un privilegio.

Voglio spendere un po' di parole per Edoardo e Lorenzo, due amici che mi hanno accompagnato e supportato in alcuni momenti di estrema difficoltà in cui mi sentivo sopraffatto da una montagna di pensieri e dubbi. Questo per me è stato un viaggio lungo, emotivamente difficile e psicologicamente insidioso, cominciato in piena pandemia, dopo una seduta di laurea svolta da casa, isolato da amici e familiari e dopo l'ennesimo trasferimento che mi ha ulteriormente allontanato da persone a me care. Poi però ho avuto l'opportunità e la fortuna di convivere con questi due ragazzi meravigliosi che reputo come fratelli. Sarò eternamente grato per tutti i momenti e le esperienze, sia positive che negative, che abbiamo condiviso durante la nostra convivenza a casa "Sasageyo". Ragazzi, per me siete una grande fonte di ispirazione per la mia crescita come persona. Edoardo ho sempre preso come riferimento la tua compostezza e il tuo modo di approcciarti alle situazioni con impegno e serietà. Di te Lorenzo ammiro la tua capacità di mettere sempre un sorriso sul volto delle persone che ti circondano con la tua simpatia e genuinità. Voglio che sappiate che il tempo passato con voi mi ha permesso di capire il tipo di persona che voglio diventare e anche se sarà difficile vi prometto che mi impegnerò per avvicinarmici il più possibile.

Voglio ringraziare anche Renata e Cristina, due amiche preziose e sempre presenti nei momenti più importanti. Renata, grazie per aver sempre fatto sentire la tua presenza anche nei periodi in cui ero assente e distaccato, per avermi spronato a uscire dalla mia bolla di pensieri e divertirmi. Ti ringrazio davvero tanto. Cristina, anche se per via della tua lontananza non abbiamo molte occasioni per passare del

tempo insieme, sappi che tengo immensamente ai momenti che abbiamo condiviso e spero che avremmo altre occasioni per viverne altri. Ragazze siete davvero fantastiche e spero che riuscirò a supportarvi nei vostri percorsi come voi avete fatto con me.

Infine, ma non meno importanti, vorrei ringraziare anche Sara, Kisela, Alessia, Anna, Gianluca e Lorenzo. Siete delle persone meravigliose e sono davvero contento di avervi conosciuto e di aver fatto amicizia con tutti voi. Ho iniziato questo percorso immerso nella solitudine e lo sto concludendo circondato da persone a cui voglio davvero bene, cosa che mai mi sarei neanche lontanamente sognato qualche anno fa.

Grazie di cuore a tutti di esserci, vi voglio bene.

Ancona, Febbraio 2024

Andrian Melnic

Sommario

Nell'attuale contesto della diagnosi medica della polmonite da COVID-19, le tecniche di imaging hanno acquisito un ruolo di primaria importanza. L'ecografia polmonare (Lung Ultrasound - LUS) in particolare, presenta una serie di vantaggi in quanto economica, sicura e non invasiva. In questo ambito, il dataset di ecografie polmonari Italian COVID-19 Lung Ultrasound (ICLUS), fornito dal Laboratorio di Ultrasonografia di Trento (ULTRa), presenta un sistema di punteggio a 4 livelli di gravità della patologia. Tramite l'utilizzo di tale dataset, questa tesi esplora l'applicabilità e l'efficacia dei modelli visuali basati su metodologie Transformer nel contesto della classificazione del grado di gravità della malattia COVID-19, confrontandoli con la Rete Neurale Convoluzionale (CNN) di riferimento, ResNet50. L'attenzione si concentra su due approcci principali: il primo consiste nell'utilizzo di un modello *pure attention* chiamato Shifted Window Transformer (Swin), basato esclusivamente su meccanismi di Multi-Head Self Attention (MSA) senza convoluzioni; il secondo prevede l'integrazione dell'MSA all'interno di una backbone convoluzionale per creare un modello ibrido chiamato Bottleneck Transformer Network (BoTNet). Lo scopo della ricerca è anche di determinare quale di questi approcci sia più efficace per il compito di classificazione e affronta le sfide poste dalla limitata disponibilità di dati nel dataset, valutando l'impiego del transfer learning come soluzione per ottimizzare l'addestramento del modello Swin e migliorare le performance.

I risultati del confronto tra i modelli addestrati da zero evidenziano che il modello ibrido BoTNet50 è in grado di superare in termini di prestazioni la ResNet50, mostrando una maggiore efficacia e consistenza nella classificazione delle immagini ecografiche, ottenendo un F1-Score di 0.6025 contro 0.5896. In confronto, il modello Swin non ha ottenuto risultati competitivi quando addestrato da zero su questo dataset, raggiungendo un F1-score di 0.4682. Tuttavia, è migliorato significativamente con l'impiego del transfer learning, riuscendo a raggiungere un F1-score di 0.6513 e superando ResNet50 su tutte le metriche. Durante lo studio di ablazione effettuato bloccando in fase di addestramento la conoscenza di particolari strati del modello Swin pre-allenato, è emerso che è possibile raggiungere prestazioni paragonabili a quelle ottenute con un modello completamente addestrato, con una complessità computazionale significativamente ridotta, anche rispetto a ResNet50. Infine, l'analisi delle Grad-CAM ha mostrato una maggiore sensibilità del Swin nel rilevare strutture complesse rispetto a ResNet50 e la capacità di porre l'attenzione a contesti più ampi, catturando le relazioni spaziali tra artefatti maggiormente diversificati e posizionati in varie aree dell'immagine.

Abstract

In the current context of COVID-19 pneumonia diagnosis, imaging techniques have become increasingly important. Lung Ultrasound (LUS), in particular, is advantageous due to its cost-effectiveness, safety, and non-invasive nature. In this setting, the Italian COVID-19 Lung Ultrasound (ICLUS) dataset, provided by the Trento Ultrasound Laboratory (ULTRa), features a four-level severity scoring system. This thesis explores the applicability and effectiveness of Transformer-based visual models in classifying COVID-19 severity levels, comparing them with the reference Convolutional Neural Network (CNN), ResNet50. The focus is on two approaches: using a pure attention model called Shifted Window Transformer (Swin), based solely on Multi-Head Self Attention (MSA) without convolutions; and creating a hybrid model named Bottleneck Transformer Network (BoTNet) by integrating MSA into a convolutional backbone. The research aims to determine which approach is more effective for classifying ICLUS ultrasound images, considering accuracy and model generalizability. It also tackles the challenges of limited data availability, evaluating the use of transfer learning to optimize Swin model training and improve its performance.

The comparison results of the models trained from scratch show that the hybrid BoTNet50 outperforms ResNet50 in terms of performance, demonstrating greater efficacy and consistency in ultrasound image classification, achieving an F1-Score of 0.6025 compared to 0.5896. In contrast, Swin did not achieve competitive results when trained from scratch on this dataset, with an F1-Score of 0.4682. However, it improved significantly with transfer learning, achieving an F1-Score of 0.6513 and surpassing ResNet50 on all metrics. An ablation study, conducted by freezing certain layers during Swin model training, revealed that performance comparable to a fully trained model can be achieved with significantly reduced computational complexity, even when compared to ResNet50. Finally, Grad-CAM analysis showed Swin's greater sensitivity in detecting complex structures than ResNet50 and its ability to focus on broader contexts, capturing spatial relationships between more diversified artifacts located in different areas of the image.

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Introduzione

1.1 Imaging Medico COVID-19

Durante il periodo pandemico di SARS-COV-2 la diagnostica tramite immagini ha giocato un ruolo chiave nel processo di diagnosi e monitoraggio della patologia COVID-19 [2]. In particolare, la TC (Tomografia Computerizzata) si è rivelata uno strumento fondamentale in quanto caratterizzata da un elevata sensibilità diagnostica e in grado di rilevare anomalie polmonari ancor prima della comparsa delle manifestazioni cliniche, permettendo così una diagnosi precoce per pazienti sospetti [3]. Tuttavia la TC presenta dei svantaggi non indifferenti. Primo fra tutti la poca disponibilità delle apparecchiature in relazione all'elevato numero di pazienti infetti e poi l'elevata rischiosità del trasporto dei pazienti nelle strutture abilitate, vista la facilità di diffusione del virus ed i lunghi tempi di sterilizzazione delle attrezzature dopo ogni esame.

L'ecografia polmonare *bedside*, al letto del malato, in un contesto clinico come quello della polmonite da COVID-19 presenta una serie innegabile di vantaggi; migliore facilità di disinfezione, minore area di contatto con i pazienti e la possibilità di eseguire l'esame senza spostare il paziente in barella, riducendo così il rischio di diffusione, pur rispettando le adeguate misure di protezione [3]. La sensibilità è minore rispetto alla TC toracica ma superiore alla radiografia per quanto riguarda le alterazioni pleuriche e sub pleuriche, tipiche della polmonite interstiziale. Tutto questo unito all'assenza di esposizioni alle radiazioni.

Le limitazioni principali della metodologia sono la scarsa capacità di rilevamento di lesioni polmonari profonde e la possibilità di diagnosticare solo le lesioni della linea pleurica, esplorabile però fino al 70% nel caso di soggetti con caratteristiche fisiche, come l'obesità, che ostacolano l'applicazione della tecnica. Da sottolineare che le lesioni riscontrate tramite l'impiego di un qualsiasi metodo di imaging (ecografie, TC o radiografie) non sono patognomoniche di COVID-19, ovvero non indicano la certezza della malattia ma in abbinamento alla storia e al quadro clinico del paziente ne rafforzano la diagnosi [2].

1.2 Ecografia Polmonare

L'ecografia polmonare (LUS - Lung Ultrasound) risulta quindi uno strumento a basso costo, veloce e poco invasivo per la diagnosi e il monitoraggio dei pazienti con sospetta o confermata infezione. La metodologia sfrutta la presenza di artefatti che derivano dall'interazione del fascio ultrasonoro con i tessuti. Le onde, con frequenza superiore ai 20 kHz (normalmente dai 3 ai 15 MHz) vengono riflesse in modo diverso dai tessuti con differente impedenza acustica, ovvero la resistenza intrinseca della materia all'attraversamento delle onde stesse [4]. L'impedenza acustica dipende dalla densità e dalla velocità di propagazione all'interno del tessuto, e determina la percentuale del fascio ultrasonoro riflesso o trasmesso al passaggio da un tessuto all'altro. Quando il fascio ultrasonoro incontra un'interfaccia acustica, cioè il confine tra due tessuti con impedenza acustica diversa, si genera un'eco, che viene ricevuta dalla sonda e viene elaborata per formare l'immagine ecografica. Il polmone normale presenta una bassa riflettività, poiché la differenza di impedenza acustica tra l'aria presente al loro interno e il tessuto connettivo è molto elevata, quindi la maggior parte delle onde sonore viene riflessa dalla superficie pleurica facendo sì che il polmone sia poco esplorabile all'eco. Le alterazioni patologiche però, aumentano la riflettività in quanto riducono la differenza di impedenza acustica tra l'aria e il tessuto connettivo permettendo alle onde sonore di penetrare più in profondità e la conseguente generazione degli echi alterati, ovvero gli **artefatti visivi** (Figura 1.1) che possono essere utili ai fini diagnostici [4]. Alcuni esempi di alterazioni tipiche del COVID-19 che possono essere rilevate osservando questi artefatti sono il versamento pleurico, la presenza di consolidazioni, edema interstiziale polmonare e il pneumotorace.

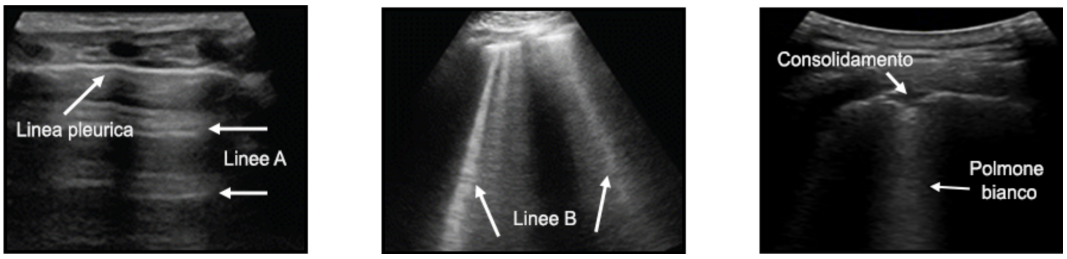


Figura 1.1: Artefatti ecografie polmonari

1.3 Problemi e sfide del dominio

L'analisi delle immagini ecografiche richiede un certo livello di esperienza e competenza da parte degli operatori risultando quindi particolarmente suscettibile all'errore umano. Inoltre, la quantificazione del grado di alterazione della pleura polmonare si basa spesso su scale soggettive e qualitative, con conseguenti limitazioni alla comparabilità e alla riproducibilità dei risultati. Per fornire supporto in fase di elaborazione e analisi delle ecografie negli ultimi anni sono state studiate varie metodologie basate

sull'apprendimento automatico, sopratutto nel campo della visione artificiale. Tra queste, le Reti Neurali Convoluzionali (CNN) si sono rivelate efficaci nella classificazione, segmentazione e localizzazione delle immagini ecografiche, sfruttando la capacità di apprendere gerarchie di filtri convoluzionali in grado di catturare caratteristiche visive rilevanti. L'adozione del Deep Learning (DL) in questo ambito sembra offrire soluzioni promettenti per superare queste difficoltà, tuttavia l'impiego di questi modelli introduce nuove sfide.

La natura stessa delle ecografie polmonari, con la loro tipica manifestazione di artefatti come le linee A e B, aree di consolidamento e la presenza di effusioni pleuriche, pone sfide significative. Queste immagini spesso mancano di chiare demarcazioni e presentano un alto grado di variabilità intra e inter-paziente, rendendo la standardizzazione dell'analisi un processo complesso. Di conseguenza, l'addestramento efficace dei modelli DL richiede un ampio dataset annotato con maggior precisione possibile e che copra un'ampia varietà di casi clinici, il che non è possibile nella maggior parte dei casi a causa di fattori quali: la stretta regolamentazione sulla privacy e la riservatezza dei dati dei pazienti; la limitatezza delle risorse per la raccolta e l'annotazione dei dati; la vasta variabilità e complessità dei casi clinici. Un altro aspetto che merita attenzione è il bilanciamento tra sensibilità e specificità nell'identificazione delle anomalie. Un modello eccessivamente sensibile potrebbe portare a un alto tasso di falsi positivi, mentre un'alta specificità potrebbe non catturare tutti i casi patologici.

Mentre le CNN eccellono nell'identificare pattern localizzati e strutture geometriche, possono avere difficoltà nel catturare relazioni spaziali complesse e caratteristiche globali nelle immagini ecografiche, un limite che diventa evidente in presenza di rumore o di strutture patologiche piccole, meno definite o estese in varie regioni dell'immagine. Recentemente, nel campo dell'elaborazione del linguaggio naturale (NLP) i modelli **Transformer** hanno raggiunto un enorme successo, diventandone lo standard *de-facto* per diversi task, come la traduzione automatica, la generazione del testo e la comprensione del linguaggio grazie proprio alla loro capacità di analizzare contesti più ampi e catturare relazioni complesse. Questo notevole successo ha spinto i ricercatori a studiare l'applicabilità del **Multi-Head Self Attention (MSA**), il meccanismo che sta alla base dei Transformer, nel campo della visione artificiale dando così origine ad una nuova serie di modelli, i quali hanno dimostrato risultati competitivi nei confronti delle CNN in diversi task. È nato quindi un fervente interesse della comunità scientifica in numerosi ambiti della computer vision. In particolare, nel settore dell'imaging medico si è assistito a una crescente adozione di metodologie basate su Transformer, con un aumento significativo dopo l'emergere dei **Vision Transformer (ViT)**. Questa tendenza è ormai dominante in prestigiose conferenze e riviste del settore tuttavia, il numero delle pubblicazioni che trattano queste metodologie applicate agli Ultrasuoni è piuttosto limitato. I ViT emergono come un'alternativa promettente alle CNN, grazie alla loro capacità di catturare relazioni a lungo raggio all'interno dell'immagine. Attraverso la MSA, sono in grado di analizzare l'intero contesto dell'immagine, superando così alcune limitazioni delle

CNN nell'identificare le relazioni spaziali complesse e relazioni globali nelle ecografie. Infine, l'addestramento dei modelli basati su Transformer nel contesto clinico presenta un problema non indifferente, ovvero che i Transformer **richiedono spesso grandi quantità di dati** per essere addestrati in modo efficace, presentando quindi, una limitazione in ambito medico dove i dataset annotati e di alta qualità non sono frequenti a causa della complessità e della variabilità dei casi clinici.

1.4 Obbiettivi

Gli obbiettivi di questa tesi sono molteplici. Il primo fra tutti è quello di esplorare il potenziale e l'efficacia delle metodologie Transformer nel contesto della classificazione di immagini ecografiche polmonari. La ricerca si propone di valutare quale tra gli approcci - puro MSA o ibrido con convoluzioni - risulta più vantaggioso in termini di performance, considerando la capacità di questi modelli di adattarsi alle sfide specifiche del dominio ecografico.

Tra queste la limitata disponibilità di dati, che potrebbe ostacolare l'efficacia dei Transformer, notoriamente esigenti in termini di volumi di dati per l'addestramento. Si indaga, quindi, se tramite il **transfer learning** si possa mitigare questa limitazione. Un altro aspetto fondamentale riguarda l'interpretazione dei risultati; nonostante la capacità del MSA di evidenziare le aree d'interesse nelle immagini, l'interpretazione di queste attenzioni in un contesto clinico resta da verificare. Le mappe di attenzione generate dai modelli vengono valutate per verificare la loro capacità di evidenziare in maniera più o meno corretta le zone interessate dagli artefatti. Infine si esaminano le performance dei modelli, sopratutto in termini di sensibilità. Tutto questo svolgendo anche un confronto con il modello CNN di riferimento.

Capitolo 2

Related Work

Questo capitolo è dedicato all'esplorazione delle metodologie di classificazione applicate sulle ecografie (LUS - Lung Ultrasound), in particolare sul dataset ICLUS. I modelli Vision Transformer (ViT) sono stati impiegati con successo nella diagnosi medica, in particolare nell'ambito dell'imaging medico. Un esempio di applicazione dei ViT è l'Enhanced Vision Transformer Model (EVTM), progettato da Xing et al. per la diagnosi automatica della polmonite attraverso l'analisi di radiografie del torace [5]. L'EVTM combina il Vision Transformer con un autocodificatore variabile per la data augmentation dei dati, migliorando così la qualità della diagnosi. In un altro studio, è stato proposto un modello ViT per la classificazione del ictus su immagini di tomografia computerizzata (TC) del cervello, ottenendo un'accuratezza fino al 98.75% [6], mentre nel ambito dell'imaging a risonanza magnetica (MRI), Tummala et al. hanno proposto un ensemble di modelli ViT per la classificazione automatizzata dei tumori cerebrali, ottenendo un'accuratezza del 98.7% [7].

Da sottolineare che sebbene nel campo dell'imaging medico le metodologie Transformer siano state applicate con successo a diverse modalità, per quel che riguarda le ecografie il numero di ricerche che sfruttano tali metodologie è limitato e, nel caso specifico del dataset ICLUS, nullo.

Nella parte iniziale di questo capitolo viene presentata una breve panoramica dello stato dell'arte della classificazione sulle ecografie ICLUS. Infine verranno presentati due lavori basati su approcci differenti nell'utilizzo dei meccanismi dei Transformer per analizzare ecografie non LUS.

2.1 Stato dell'Arte della Classificazione su ICLUS

Nello stato dell'arte della classificazione sul dataset ICLUS, diverse ricerche hanno fornito contributi significativi. Roy et al. hanno presentato in [8] un architettura che combina una Spatial Transformer Network (STN), una rete convoluzionale (CNN) e un approccio ordinale per classificare le ecografie del ICLUS a livello di fotogramma in base alla gravità (Figura 2.1 e 2.2). Un contributo secondario, ma non meno importante, di questo studio è lo sviluppo di un metodo basato su uninorms per l'aggregazione delle previsioni a livello di singoli frame e per la determinazione del punteggio corrispondente al video.

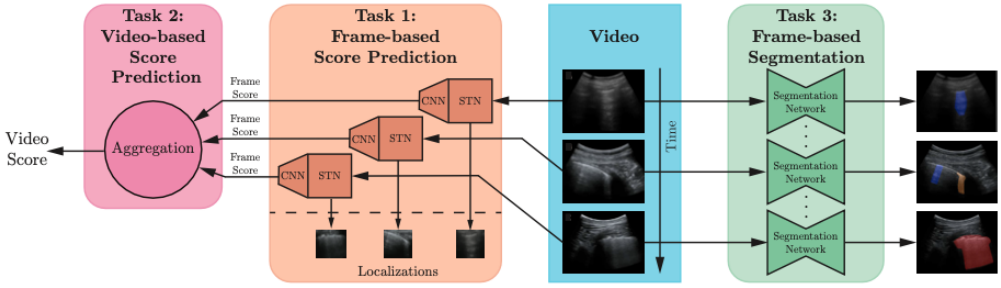


Figura 2.1: Framework Roy et al.

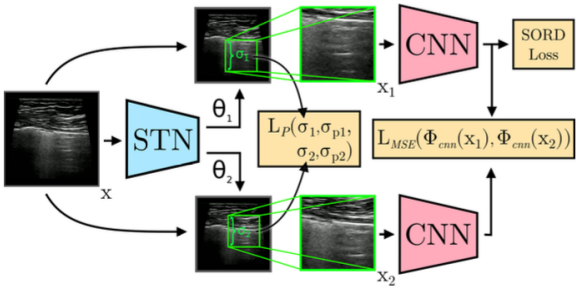


Figura 2.2: Architettura Reg-STN

Il lavoro [9] di Oz Frank et al. propone un framework per addestrare reti neurali deep con particolare focus sull'interpretabilità. Durante l'addestramento ai frame LUS viene integrata conoscenza del dominio sotto forma di due canali aggiuntivi contenenti artefatti verticali individuati automaticamente e una maschera di distanza dalla linea pleurica per migliorare le prestazioni dei modelli sotto analisi (Figura 2.3).

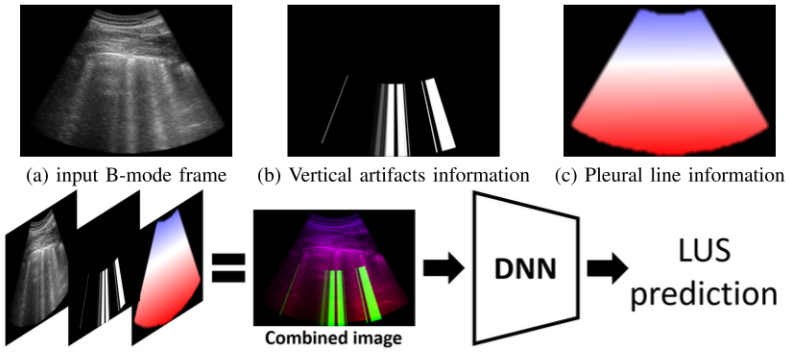


Figura 2.3: Integrazione della conoscenza del dominio in Frank et al.

Khan et al. in [10] presentano un approccio metodologico di riferimento per l'applicazione dell'intelligenza artificiale ai dati di ecografia polmonare LUS che prevede l'analisi dei dati a livello di frame, video e prognosi e presenta risultati promettenti per la diagnosi assistita di COVID-19.

2.2 Metodologie basate sui Transformer applicate alle ecografie

In [11] Płotka et al. introducono **BabyNet**, un'architettura ibrida ispirata al Bottleneck Transformer [1] per l'analisi di ecografie fetali a livello video. BabyNet (Figura 2.4) ha come base una 3DResNet18 [12] modificata sostituendo gli ultimi 2 blocchi residuali con 2 Residual Transformer Module (RTM). Questi moduli sono caratterizzati dalla presenza del 3D-MSA, un Multi-Head Self Attention come descritto in 2.4.4 in grado di elaborare anche informazioni temporali oltre a quelle spaziali. Il lavoro introduce anche codifica posizionale relativa (RPE) spazio-temporale. Quest'ultima è una codifca posizionale alternativa a quella assoluta presentata in 2.4.1 che prende come riferimento il Bias Posizionale Relativo presentato da Shaw et al. [13] a cui viene aggiunta anche una codifica temporale per aiutare il modello a capire l'ordine dei fotogrammi.

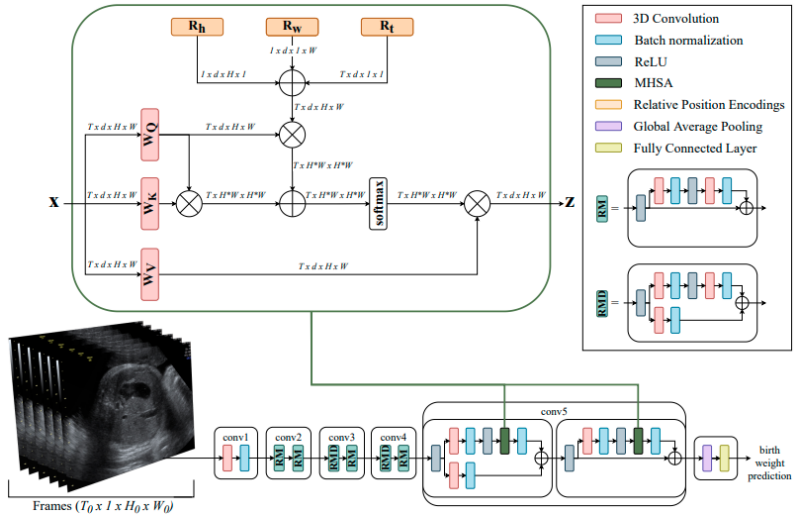


Figura 2.4: Architettura BabyNet

Recentemente [14] ha esplorato il potenziale dei modelli Vision Transformer (ViT) per la classificazione delle immagini di ultrasuoni del seno, confrontandoli con le reti neurali convoluzionali. Nello studio sono stati utilizzati modelli pre addestrati per applicare la tecnica del Transfer Learning in modo da adattare i modelli pre-addestrati a un compito di classificazione a tre classi (benigno, maligno o normale) su due diversi set di dati di ultrasuoni del seno (BUSI e B). I modelli ViT hanno mostrato performance comparabili o anche migliori delle CNN dimostrando il potenziale dei modelli ViT nei task di classificazione di immagini ecografiche se pre-allenati su dataset più grandi.

2.3 Self Attention per la classificazione LUS

Il successo del Transformer nel campo dell'elaborazione del linguaggio naturale ha suscitato un crescente interesse nel cercare di sfruttare le sue proprietà anche nella computer vision. Nell'imaging medico, e in particolare nella diagnostica ecografica, le immagini si caratterizzano per una complessità intrinseca e un'elevata densità di strutture interrelate. Un esempio specifico nel nostro caso può essere la presenza di artefatti verticali *B-line* e consolidamenti sub pleurici, che si manifestano a causa di interruzioni della linea pleurica. In questo contesto la Self Attention, approccio che viene discusso più dettagliatamente nella sezione 2.4.3, mostra un potenziale non indifferente in quanto la capacità di gestire le dipendenze a lungo raggio tra regioni, anche distanti, nell'immagine ecografica potrebbe offrire un'analisi più approfondita e accurata di tali strutture. Inoltre eseguendo il calcolo dell'attenzione utilizzando più teste tramite il meccanismo di Multi-Head Self Attention (MSA) (Sezione 2.4.4) si potrebbero generare più rappresentazioni simultaneamente migliorando la capacità generalizzativa del modello.

In questa tesi, sono stati sperimentati due approcci per l'uso della self attention nella classificazione di ecografie polmonari (LUS):

1. Il primo si basa sull'utilizzo di un modello basato interamente sulla MSA senza l'impiego di convoluzioni, quindi un modello *pure attention*. Questa categoria di modelli stanno ottenendo popolarità nella classificazione delle immagini grazie alla loro capacità di catturare relazioni a lungo raggio tra i pixel. L'utilizzo di self-attention senza convoluzioni consente di gestire efficacemente le relazioni spaziali complesse all'interno delle immagini, evitando le limitazioni delle convoluzioni, come la dimensione fissa dei kernel e la focalizzazione su caratteristiche locali. Lo svantaggio è la necessità di una maggiore quantità di dati per l'addestramento dei moduli MSA in quanto la mancanza di convoluzioni porta l'assenza di determinate proprietà intrinseche delle convoluzioni stesse, le quali nelle CNN introducono delle conoscenze pregresse consentendo di catturare determinati pattern con maggiore facilità;

2. Il secondo approccio cerca di combinare invece il meccanismo di MSA con una backbone convoluzionale, creando un modello ibrido. La parte convoluzionale funge da estrattore di feature map astratte a bassa risoluzione che conservano parte di queste proprietà. Successivamente viene impiegata la MSA globale per processare e aggregare le informazioni contenute nelle feature map convoluzionali [15]. Questa tipologia di architettura permette al modello di catturare efficacemente sia le caratteristiche discriminative globali che locali, portando a un miglioramento nelle prestazioni nei compiti di classificazione senza accusare il problema di efficienza dei dati dei modelli pure attention.

2.4 Vision Transformer

Mentre il Transformer è diventato lo standard *de-facto* per task di elaborazione del linguaggio naturale, fino all'avvento del Vision Transformer la tecnica dell'attenzione veniva applicata solo in combinazione con le reti convoluzionali, oppure sostituendo parti delle reti convoluzionali mantenendo però la loro struttura invariata [16]. Il Vision Transformer (ViT) introdotto da Dosovitskiy et al. [16] è il primo modello a sfruttare solamente la Self Attention (SA) e il Multi-Head Self Attention (MSA) [17]. Gran parte dell'architettura rimane invariata rispetto al Transformer NLP. La differenza principale sta nella modalità con cui l'input viene gestito, essendo ovviamente composto da immagini e non sequenze testuali, e l'assenza del decodificatore, non necessario visto che non è prevista la generazione di alcun tipo di output. L'architettura presenta però delle problematiche, che verranno descritte successivamente, per le quali è stata reputata non ideale all'utilizzo sul datataset ICLUS. Tuttavia, è importante sottolineare che molti dei concetti del Vision Transformer stanno alla base delle architetture analizzate in questa tesi. Lo scopo di questa sezione e quindi, quello di fornire una breve e doverosa panoramica introduttiva delle componenti principali del ViT per poi passare, nel capitolo successivo, alle architetture utilizzate.

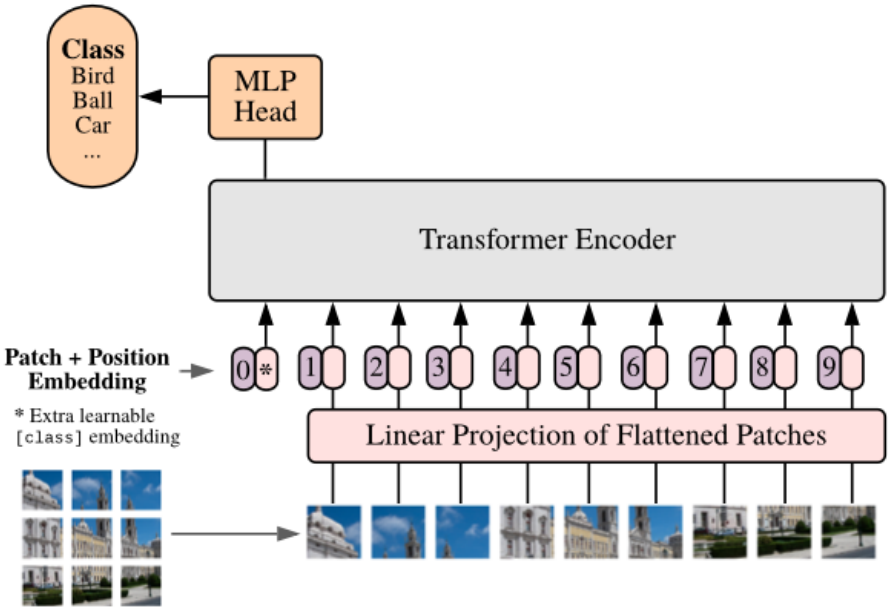


Figura 2.5: Vision Transformer

2.4.1 Input Embedding e Codifica Posizionale

Come si può vedere nella Figura 2.5, il ViT inizia dividendo l'immagine di input in una griglia di patch quadrate di dimensione fissa (ad esempio, 16x16 pixel) non sovrapposte. Le patch poi vengono appiattite e trasformate in vettori unidimensionali che poi vengono passati come input a una rete completamente connessa che applica

una trasformazione lineare per ottenere una sequenza di vettori della dimensione desiderata d_{model}. Vista l'importanza della posizione delle patch all'interno dell'immagine, ai vettori viene aggiunta una codifica che da informazioni sulle posizioni delle patch all'interno dell'immagine. Ciò è necessario vista l'assenza di convoluzioni e ricorrenze [17] e le modalità di codifica posizionale variano in base all'implementazione. La sequenza di vettori così ottenuta è l'input del codificatore.

2.4.2 Codificatore

La struttura del codificatore (Figura 2.6) è composta da una pila di N layer identici, ognuno dei quali contiene due moduli principali: il primo è il **Multi-Head Self Attention (MSA)**, che riceve come input le query, le chiavi e i valori e calcola i punteggi di attenzione tra gli elementi della sequenza in input; il secondo è una rete Feed-Forward completamente connessa, che applica una trasformazione non lineare ai vettori di output del MSA. Dopo ogni modulo è presente una connessione residuale seguita da una Layer Normalization [18] che rispettivamente aiutano a evitare la dispersione del gradiente e a stabilizzare e velocizzare il processo di addestramento [17].

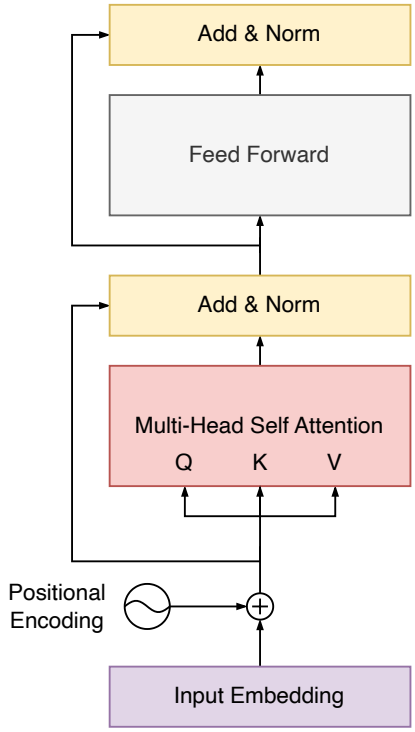


Figura 2.6: Codificatore del Transformer

2.4.3 Scaled Dot-Product Attention (SA)

L'attenzione può essere vista come un'operazione che mappa una query e un insieme di coppie chiavi-valori su un output, calcolato come una somma pesata dei

valori. Il peso assegnato a ciascun valore è ottenuto in funzione alla query e alla chiave corrispondente. Nel codificatore, l'attenzione viene calcolata tra le patch della stessa immagine. Le rappresentazioni vettoriali delle patch vengono moltiplicate per le matrici W^Q, W^K, W^V apprese durante l'addestramento per ottenere le query $Q \in \mathbb{R}^{n \times d_k}$, le chiavi $K \in \mathbb{R}^{m \times d_k}$ e i valori $V \in \mathbb{R}^{m \times d_v}$ [17].

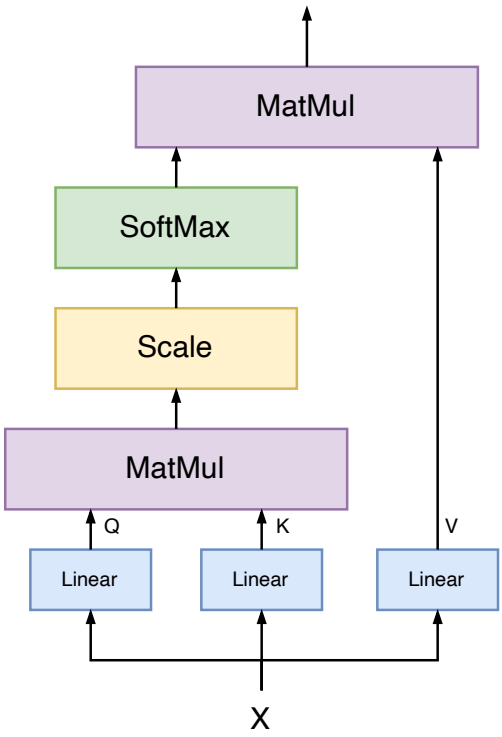


Figura 2.7: Operazione Scaled Dot-Product Attention

Il calcolo del punteggio di attenzione avviene simultaneamente per ciascuna coppia di vettori ed è espresso come segue:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.1}$$

La funzione, chiamata dagli autori *Scaled Dot-Product Attention* [17] (Figura 2.7), calcola i prodotti scalari tra le query Q e tutte le chiavi K e poi li divide per il termine $\sqrt{d_k}$ così da evitare che si generino gradienti eccessivamente piccoli per dimensioni d_k maggiori. Ai valori così ottenuti viene applicata la softmax e utilizzati come pesi per i vettori di V.

2.4.4 Multi-Head Self Attention (MSA)

La *Multi-Head Self Attention* (Figura 2.8) è un'estensione della Scaled Dot-Product Attention che permette di sfruttare più spazi di rappresentazione in parallelo. I vettori di Q, K, V vengono proiettati in h sottospazi più piccoli di dimensione d_{model}/h, dove

h è il numero di teste ed è un iperparametro del modello [17]. Ogni testa ha un proprio insieme di matrici W_i^Q, W_i^K, W_i^V che vengono aggiornate durante l'addestramento. Una volta ottenuti, gli output di tutte le teste vengono poi concatenati e moltiplicati per una matrice di proiezione finale W^O, che riporta la dimensione a d_{model}.

$$\text{MultiHead} = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2.2}$$

$$\text{dove} \quad \text{head}_i = \text{Attention}(XW_i^Q, XW_i^K, XW_i^V) \tag{2.3}$$

Dove le proiezioni sono matrici parametrizzate $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ e $W^O \in \mathbb{R}^{hd_v \times d_{model}}$ con $d_k = d_v = d_{model}/h$. Questo meccanismo

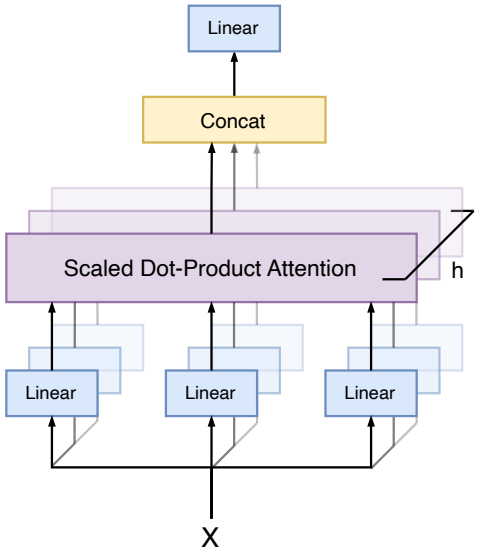


Figura 2.8: Modulo Multi-Head Self Attention

permette al modello di apprendere contemporaneamente più relazioni tra le patch a diversi livelli di astrazione, con un costo computazionale similare a quello di una Scaled Dot-Product Attention su singola testa con dimensionalità d_{model}.

A differenza di ciò che avviene nelle Reti Neurali Convoluzionali (CNN), dove dall'immagine vengono generate feature map locali tramite convoluzioni e pooling, i ViT elaborano l'attenzione sull'intera immagine offrendo una visione olistica del campo visivo. Nelle CNN, infatti, il processo di convoluzione utilizza kernel di dimensioni fisse per estrarre feature locali dall'immagine. Ogni kernel si concentra su una specifica regione, limitando l'analisi alle informazioni spaziali immediate. Inoltre, il pooling riduce ulteriormente le dimensioni spaziali delle feature map, enfatizzando le caratteristiche dominanti ma potenzialmente perdendo altri dettagli. I ViT, invece, impiegano meccanismi di attenzione che assegnano pesi a ciascuna parte dell'immagine, creando una mappa di attenzione (attention map). Questa mappa evidenzia le aree di maggiore rilevanza, permettendo al modello di considerare le relazioni a lungo raggio tra diverse regioni dell'immagine e grazie all'MSA i ViT

sono in grado di calcolare varie attention map in parallelo.

2.4.5 Problemi del ViT

Il ViT presenta delle limitazioni non indifferenti. Prima fra tutte è la dimensione delle patch che poco si adatta alla complessità delle ecografie che possono presentare una vasta gamma di caratteristiche visive, come variazioni di texture, artefatti e strutture di dimensioni diverse. Le patch di dimensioni troppo grandi potrebbero negare al modello la capacità di catturare questi dettagli fini. Di conseguenza, la rigidità nella dimensione delle patch può portare a una perdita di informazioni cruciali o a una comprensione parziale delle immagini ecografiche rendendo il modello inadatto ai task di *dense prediction* come la segmentazione semantica. Inoltre, l'architettura non scala in modo efficiente al diminuire della dimensione delle patch o all'aumentare della risoluzione dell'immagine in quanto il calcolo globale della MSA ha complessità computazionale quadratica rispetto al numero delle patch. Per questi motivi, per la classificazione delle ecografie LUS, si è deciso di utilizzare in alternativa il modello Swin Transformer che, come viene mostrato in 3.2, presenta delle caratteristiche molto interessanti che lo rendono un buon candidato come backbone general purpose per la computer vision [19].

Capitolo 3

Metodologie

3.1 Dataset ICLUS

3.1.1 Introduzione

Il dataset **Italian COVID-19 Lung Ultrasound (ICLUS)** [20, 21, 8] è stato rilasciato dal Laboratorio di Ultrasonografia di Trento (ULTRa), Dipartimento di Ingegneria dell'Informazione e Scienze Informatiche, Università di Trento. Comprende 277 video di ecografia polmonare, per un totale di 58924 frame. I dati provengono da sonde lineari (13364 frame) e sonde convesse (45560 frame) e sono stati acquisiti in vari ospedali italiani, tra cui:

- BresciaMed a Brescia,
- l'Ospedale Generale Valle del Serchio a Lucca,
- la Fondazione Policlinico Universitario A. Gemelli IRCCS a Roma,
- l'Ospedale Generale di Tione (TN),
- la Fondazione Policlinico Universitario San Matteo IRCCS a Pavia.

La Figura 3.1 mostra la panoramica del dataset ICLUS.

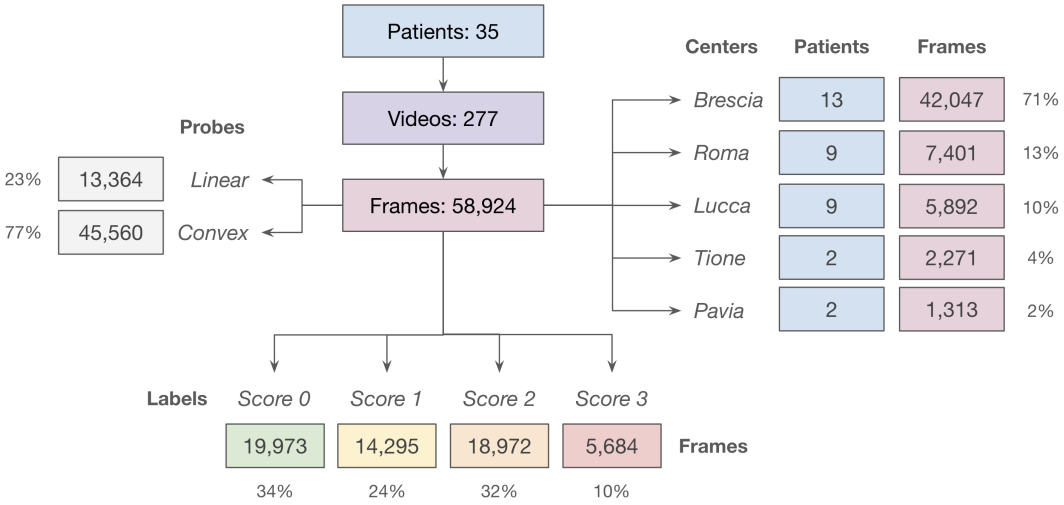


Figura 3.1: Panoramica del dataset ICLUS

3.1.2 Processo di annotazione

Il processo di annotazione di ICLUS è stato stratificato in quattro livelli distinti per garantire l'oggettività e la precisione delle annotazioni. Inizialmente, il punteggio è stato assegnato frame per frame da quattro studenti magistrali con conoscenze specifiche nelle ecografie. Successivamente, queste annotazioni sono state validate da uno studente di dottorato esperto in ecografie polmonari **Lung Ultrasound (LUS)**. Il terzo livello di validazione è stato effettuato da un ingegnere biomedico con oltre dieci anni di esperienza in LUS e infine, il quarto è stato condotto da medici con più di dieci anni di esperienza in LUS con un accordo tra gli operatori del 67%.

3.1.3 Sistema di valutazione

Ogni video nel dataset è etichettato secondo il sistema di punteggio descritto di seguito:

- **Score 0**: Indica la normalità, con linee pleuriche continue e regolari. Non è parte di alcun sottogruppo, poiché rappresenta l'assenza di anomalie significative. Sono presenti artefatti orizzontali (linea A). Circa il 34% dei frame corrispondono a questo punteggio.
- **Score 1**: Include casi più lievi, come indentazioni della linea pleurica con aree verticali bianche. È il gruppo base, con il minor grado di gravità tra i punteggi che indicano anomalie. Circa il 24% dei frame corrispondono a questo punteggio.
- **Score 2**: Comprende casi con linee pleuriche interrotte e aree consolidate piccole o grandi (aree più scure) con aree associate di bianco sotto l'area consolidata (polmone bianco). Circa il 32% dei frame corrispondono a questo punteggio. Questo gruppo è un sottogruppo di Score 1, poiché ogni frame con Score 2 soddisfa anche i criteri per Score 1.
- **Score 3**: Rappresenta il gruppo più grave. Include casi in cui l'area esaminata mostra un polmone bianco denso ed esteso, con o senza consolidazioni maggiori (per almeno il 50% della linea pleurica). Questo gruppo è un sottogruppo di Score 2, poiché ogni frame con Score 3 soddisfa anche i criteri per Score 2 e Score 1. Circa il 10% dei frame corrispondono a questo punteggio.

Le Figure 3.2 e 3.3 mostrano rispettivamente un esempio per ciascuna classe e la distribuzione dei frame per ogni punteggio nel dataset.

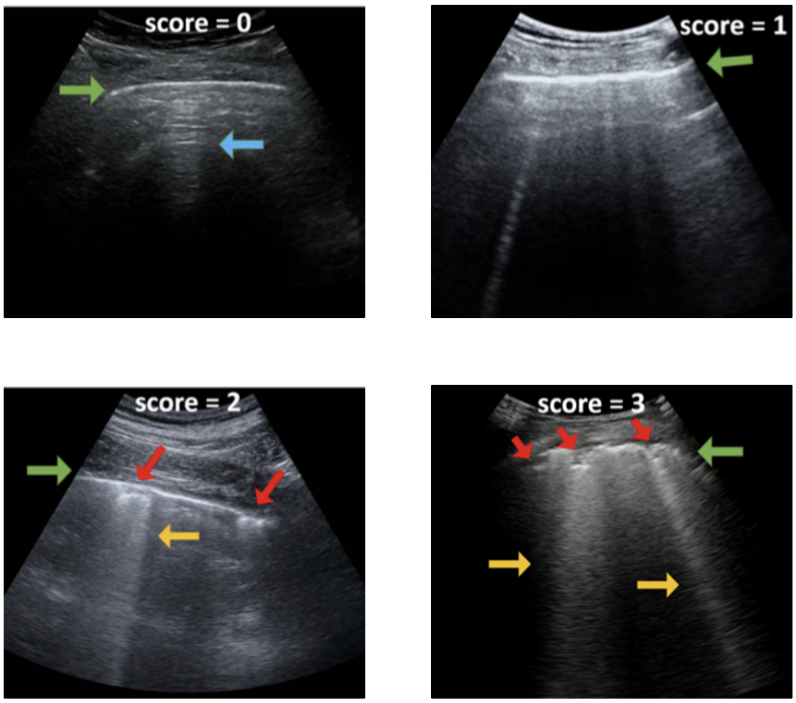


Figura 3.2: Gradi di gravità COVID-19

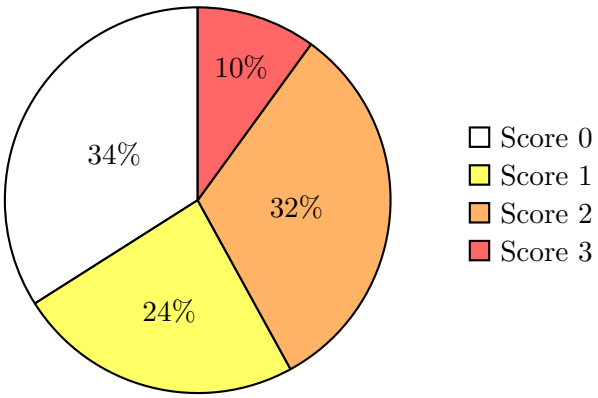


Figura 3.3: Distribuzione percentuale dei frame per ogni punteggio nel dataset

3.1.4 Preparazione del dataset

Il dataset è gestito usando un file in formato `HDF5`, ampiamente utilizzato per l'archiviazione di dataset di grandi dimensioni. L'accesso ai frame è gestito in maniera ottimizzata via codice utilizzando librerie dedicate. Per facilitare l'accesso e l'analisi specifica di ciascun frame, permettendo una gestione più efficiente e organizzata, è stato creato un sistema di mappatura indicizzato che associa ogni frame ai rispettivi gruppi e video presenti nel dataset.

Per quanto riguarda la suddivisione dei frame, si è optato per la tecnica hold-out a tre set comprendendo train, validazione e test, nelle percentuali $60\%, 20\%, 20\%$. È importante sottolineare che la suddivisione non avviene ne a livello dei frame ne a livello dei video, ma livello di paziente. Questo perché nel dataset, per alcuni dei pazienti, sono presenti più video ecografici. Per questo bisogna far sì che i video dello stesso paziente **non finiscano** in due set differenti portando a un problema di *information leakage* tra i dati di train e test. Applicando la divisione sui pazienti evita che questo fenomeno si verifichi. Infine ci si è assicurati che ciascun set abbia la stessa distribuzione delle classi, mostrata in Figura 3.3, del dataset completo.

3.1.5 Frame preprocessing

Durante l'esecuzione del framework, tutti i frame di ciascun set (train, validation e test) vengono ridimensionati a una risoluzione standard di 224×224 pixel, per garantire uniformità e facilitare l'elaborazione da parte dei modelli di deep learning. Questo passaggio è cruciale per l'adattamento delle immagini a modelli pre-addestrati che richiedono input di dimensioni specifiche e che vedremo in seguito. Dopo il ridimensionamento avviene una fase di conversione e normalizzazione. I frame vengono normalizzati utilizzando i valori delle medie e deviazioni standard pre-calcolati relativi a ciascun canale delle immagini del train set. La normalizzazione è un passaggio essenziale per standardizzare i dati di input e facilitare la convergenza durante l'addestramento e va effettuata utilizzando la media e la deviazione standard calcolate a partire soltanto dal train set per evitare d'introdurre un bias durante l'addestramento.

3.2 Swin Transformer - Modello Pure Attention

Il Swin Transformer (**S**hifted **Win**dows) introdotto da Liu et al. [19] rappresenta un avanzamento significativo nel ambito dei Transformer per la visione artificiale, distinguendosi per il suo approccio gerarchico nell'elaborazione delle immagini. L'architettura affronta alcune limitazioni chiave dei ViT, introducendo un approccio gerarchico più efficiente e scalabile basato su MSA calcolata su finestre scorrevoli. Mentre il ViT ha aperto la strada nell'applicazione dei Transformer all'analisi delle immagini, lo Swin Transformer ha portato miglioramenti specifici in termini di complessità computazionale, gestione delle dipendenze spaziali e offre una notevole flessibilità d'utilizzo.

Il Swin Transformer si è affermato come modello innovativo nel panorama dei Transformer per la visione artificiale, emergendo come una backbone versatile ed efficace in grado di ottenere notevoli prestazioni in una varietà di compiti, come la classificazione delle immagini, il rilevamento degli oggetti e la segmentazione semantica, stabilendo nuovi standard di riferimento in questi ambiti. Proprio per queste sue proprietà, lo Swin Transformer è stato scelto come modello *pure attention*

da analizzare nel contesto del dataset ICLUS. In questa sezione viene fornita un'esposizione dettagliata dello Swin Transformer, esplorando in maniera approfondita le sue peculiari caratteristiche e i meccanismi operativi.

3.2.1 Architettura e Patch Merging

La Figura 3.4 presenta una panoramica dell'architettura della versione tiny del Swin Transformer. L'immagine RGB in input viene inizialmente processata attraverso un modulo di partizionamento che la suddivide in patch non sovrapposte. Ogni patch è considerata come un token e le sue feature sono date dalla concatenazione dei valori RGB dei pixel. Ogni patch ha dimensione 4×4 pixel, la dimensione della feature per ogni patch è $4 \times 4 \times 3 = 48$.

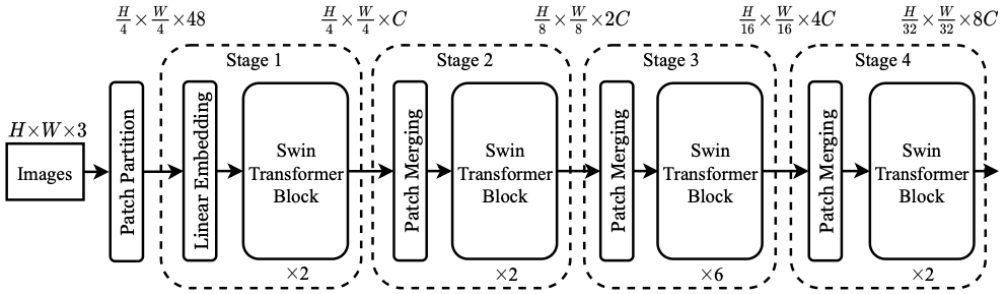
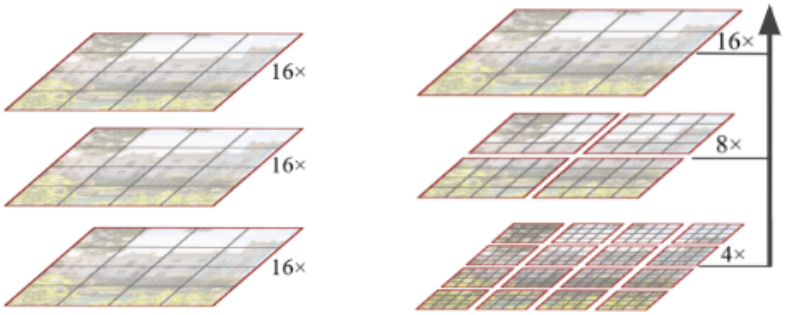


Figura 3.4: Architettura Swin Transformer

Come visto in precedenza 2.4.1 si applica una trasformazione lineare per proiettare le feature su una dimensione C. Questi "token" delle patch sono forniti come input a una coppia di **Swin Transformer Block** completando lo **Stage 1** [19]. Da notare già da subito una differenza significativa rispetto al ViT, ovvero la dimensione delle patch, 4×4 contro 16×16. Questo perché, per produrre una rappresentazione gerarchica, il numero di patch viene ridotto attraverso una fusione delle patch adiacenti. Questa fusione viene operata sui token dai **Merge Layer** presenti prima dei Swin Transformer Block a partire dal Stage 2. Il primo Merge Layer concatena quindi le feature di ciascun gruppo 2×2 di patch vicine. Questo riduce il numero di token per un multiplo di 4 e raddoppia la dimensione delle feature. In seguito alla fusione, ogni token rappresenta ora una patch di dimensione 8×8. Tale operazione di fusione si ripete a ogni successivo Stage, con i Merge Layer che continuano a ridurre il numero dei token e aumentare la dimensione delle loro feature, seguendo il principio di costruzione gerarchica delle rappresentazioni visive. Quindi, man mano che l'input attraversa i vari Stage, la dimensione spaziale dei token è ridotta, mentre quella delle loro feature aumenta. Questo processo porta alla formazione di una piramide di feature, dove ogni livello della piramide rappresenta l'input a una risoluzione spaziale più bassa ma con feature di maggiore dimensione.



(a) ViT: patch di dimensione fissa e MSA globale

(b) Swin: patch di dimensione variabile e MSA locale

Figura 3.5: Partizionanento in patch e MSA (in rosso) nel ViT e Swin Transformer

3.2.2 Swin Transformer Block e Windowed MSA

Il **Swin Transformer Block**, mostrato nella Figura 3.6 è il dedicato al meccanismo MSA nel Swin Transformer. Il blocco presenta poche variazioni strutturali rispetto al encoder del ViT. La differenza sostanziale sta nel come viene applicata la MSA. Rispetto a come avviene nel ViT, dove l'attenzione è calcolata globalmente su tutte le patch, nel Swin Transformer la MSA è confinata all'interno di piccole finestre di patch non sovrapposte (Figura 3.5). La MSA applicata in questa modalità è chiamata **Windowed MSA (W-MSA)** e riduce notevolmente la complessità computazionale dell'attenzione che da $\mathcal{O}(HW^2)$ passa a $\mathcal{O}(M^2)$, dove HW è il numero di patch e M è la dimensione della finestra. La complessità della W-MSA cresce linearmente quando M è fisso. Il calcolo della SA, rispetto a come definito in 2.4.3, subisce una leggera variazione nel utilizzo del bias relativo posizionale che viene descritta in 3.2.5. La Layer Normalization viene applicata prima dei moduli W-MSA e MLP. La connessione residuale avviene invece dopo ciascun modulo.

La W-MSA comporta però una limitazione in quanto le patch di una finestra non possono interagire con le patch di altre finestre, limitando la capacità del modello di apprendere relazioni a lungo raggio. Per questo motivo in ogni Stage, gli Swin Transformer Block sono disposti a coppie in maniera consecutiva. In questa configurazione, il primo blocco applica la W-MSA mentre quello successivo adotta una configurazione di finestre spostata di $(\frac{M}{2}, \frac{M}{2})$ pixel per garantire connessioni tra finestre adiacenti. Questa operazione è chiamata **Shifted Windowed MSA (SW-MSA)**. L'output dei due blocchi così configurati e definito come segue:

$$\hat{z}^{(l)} = \text{W-MSA}(\text{LN}(z^{(l-1)})) + z^{(l-1)} \tag{3.1}$$

$$z^{(l)} = \text{MLP}(\text{LN}(\hat{z}^{(l)})) + \hat{z}^{(l)} \tag{3.2}$$

$$\hat{z}^{(l+1)} = \text{SW-MSA}(\text{LN}(z^{(l)})) + z^{(l)} \tag{3.3}$$

$$z^{(l+1)} = \text{MLP}(\text{LN}(\hat{z}^{(l+1)})) + \hat{z}^{(l+1)} \tag{3.4}$$

dove $\hat{z}^{(l)}$ e $z^{(l)}$ denotano rispettivamente le feature in uscita dal modulo (S)W-MSA

e dal modulo MLP per il blocco l.

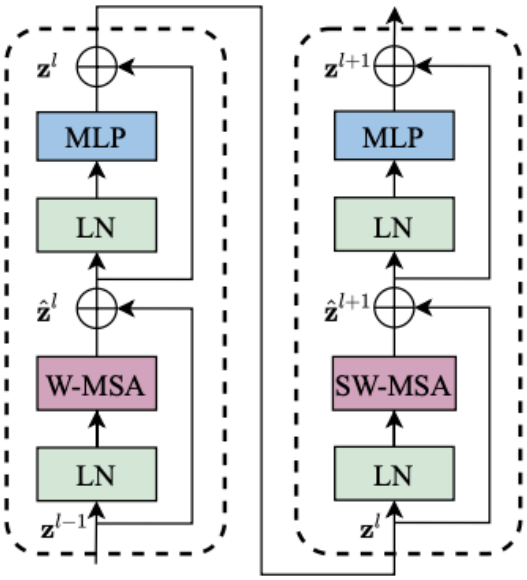


Figura 3.6: Due Swin Transforme Block consecutivi

3.2.3 Cyclic Shift

Un problema della SW-MSA è il modo in cui avviene il partizionamento delle finestre spostate, che potenzialmente può portare a un numero maggiore di finestre dove alcune finestre potrebbero risultare più piccole di $M \times M$ come mostrato in figura, 3.7.

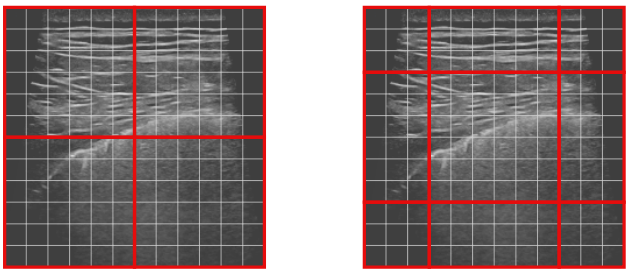


Figura 3.7: Finestre MSA (in rosso) prima e dopo lo spostamento

Il **cycle shift** che serve a gestire in modo efficiente la configurazione delle finestre spostate (SW) a causa delle finestre incomplete (più piccole di $M \times M$) che si generano dopo lo spostamento. La soluzione è di mantenere la stessa configurazione delle finestre prima dello shift e utilizzare delle sotto finestre spostandole come mostrato in Figura 3.8. Nella pratica, dopo lo spostamento, una finestra MSA può interessare diverse sotto finestre che non sono adiacenti nella feature map originale. Per mantenere l'efficienza e la correttezza del calcolo della SA all interno

delle finestre, si introduce una maschera che invece rispecchia la configurazione Shifted Window (SW). Questa operazione viene chiamata **Masked MSA** e serve a garantire che l'attenzione venga calcolata solo tra le patch che effettivamente sono vicine e fanno parte della stessa finestra nella configurazione SW [19]. Il vantaggio principale del cyclic shift è che consente di mantenere inalterato il numero di finestre rispetto al partizionamento iniziale delle finestre, evitando l'aumento della complessità computazionale che deriverebbe dal dover gestire un numero maggiore di finestre più piccole.

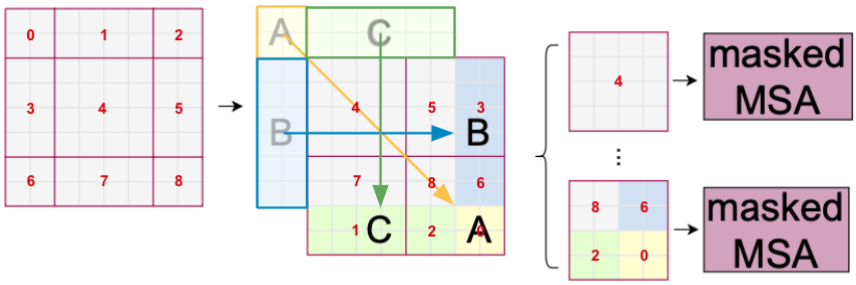


Figura 3.8: Approccio cyclic shift per la gestione efficiente della SW-MSA tramite masked MSA

3.2.4 Confronto della Complessità: MSA vs W-MSA

L'introduzione del Window-based Multi-head Self-Attention (W-MSA) offre una soluzione efficace al problema della complessità computazionale elevata associata al tradizionale Multi-head Self-Attention (MSA). Nel contesto del MSA, la complessità computazionale per un'immagine di dimensione $H \times W$ e con un numero C di canali è data da $O(HWC^2 + (HW)^2C)$. Questa formula evidenzia come la complessità sia bipartita: una parte dipendente dalla trasformazione lineare dei canali (HWC^2) e l'altra dalla matrice di attenzione che scala quadraticamente con le dimensioni dell'immagine ($(HW)^2C$).

In contrasto, l'approccio W-MSA divide l'immagine in finestre più piccole di dimensione $M \times M$ e applica l'attenzione all'interno di queste finestre locali. Di conseguenza, la complessità diventa $O(HWC^2 + M^2HWC)$. Il vantaggio chiave qui è che il termine dominante della complessità, che ora dipende dalla dimensione della finestra M piuttosto che dall'intera dimensione dell'immagine, diventa significativamente più gestibile, specialmente per immagini di grandi dimensioni. Inoltre, poiché M è tipicamente molto più piccolo di H e W, il calcolo dell'attività diventa notevolmente più efficiente.

3.2.5 Bias Posizionale Relativo

Come avviene nel ViT, anche qui è utilizzata una codifica per incorporare informazioni posizionali sulle patch. Tuttavia, mentre nel ViT questa codifica è assoluta e

viene aggiunta agli input embedding prima del codificatore [16], nel Swin Transformer si utilizza un bias posizionale relativo [22]. La codifica posizionale assoluta assegna un vettore unico a ogni posizione nella sequenza, indipendentemente dal contesto in cui compare mentre il bias posizionale relativo tiene conto della posizione delle patch rispetto alle altre patch all'interno della stessa finestra di dimensione $M \times M$.

Il bias viene introdotto nel calcolo dell'SA e si aggiunge al termine di similarità:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V,$$

dove Q, K, V sono le matrici delle query, chiavi e valori, e d è la dimensione delle chiavi/query. Poiché la posizione relativa lungo ogni asse è compresa nell'intervallo $[-M+1, M-1]$, viene utilizzata una matrice parametrizzata più piccola $\hat{B} \in \mathbb{R}^{(2M-1)\times(2M-1)}$ dalla quale vengono presi i valori di B [19].

Gli autori hanno rilevato miglioramenti significativi nell'impiego di questo bias, in confronto alle alternative che non lo adottano o che utilizzano una codifica posizionale assoluta, anche quando questa è combinata con il bias stesso [19].

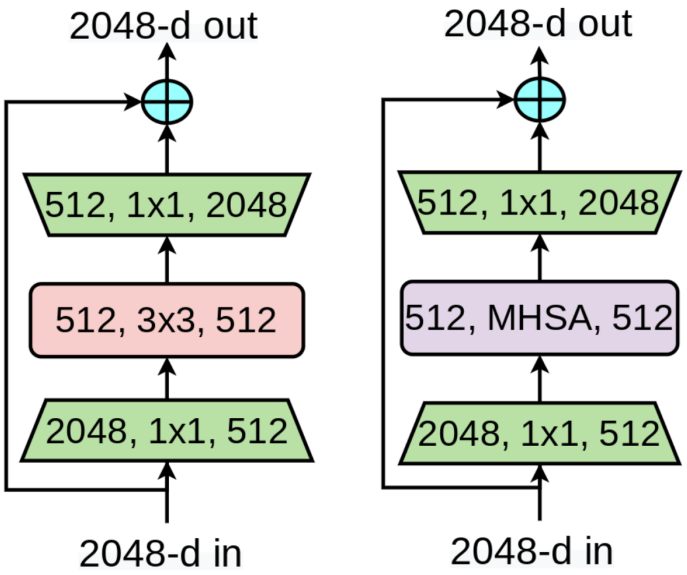
3.3 BotNeT50 - Modello Ibrido CNN e Self Attention

Il Bottleneck Transformer Network [1] o BoTNet rappresenta un approccio ibrido che combina una ResNet [23] con il meccanismo di self attention. I tradizionali blocchi Bottleneck sono sostituiti con blocchi **Bottleneck Transformer (BoT)**. Per il dataset ICLUS è stata scelta la versione basata su ResNet-50, quindi BoTNet-50, in modo da avere un confronto diretto con la stessa ResNet-50 e con lo Swin Transformer Tiny. Infatti, queste tre architetture hanno una quantità simile di parametri.

3.3.1 Blocco Bottleneck Transformer

La differenza tra le due tipologie di blocchi, come mostrato nella Figura 3.9, sta nella sostituzione della convoluzione 3×3 con un modulo MSA [1]. Nel BoTNet, a differenza del Swin Transformer, la MSA viene applicata globalmente come avviene nel ViT (Figura 3.10). La complessità quindi scala in maniera quadratica rispetto alla dimensionalità spaziale. Per questo motivo, i BoT vengono utilizzati solo nei layer più profondi della rete. La configurazione dei layer del BoTNet50 è rappresentata nella tabella 3.1.

La principale differenza, come si può osservare, è la sostituzione dei ResNet Bottleneck nell'ultimo **Stage C5** con i BoT. In questo caso la MSA globale non viene calcolata direttamente sulle immagini ma sulle feature map provenienti dai layer convoluzionali superiori che hanno una dimensionalità spaziale ridotta. Con questo design ibrido le due primitive, convoluzioni e MSA, vengono utilizzate insieme integrando in modo efficiente la capacità di apprendere relazioni complesse a lungo raggio.



(a) Bottleneck ResNet50 (b) Bottleneck Transformer

Figura 3.9: Confronto tra i blocchi ResNet Bottleneck e Bottleneck Transformer

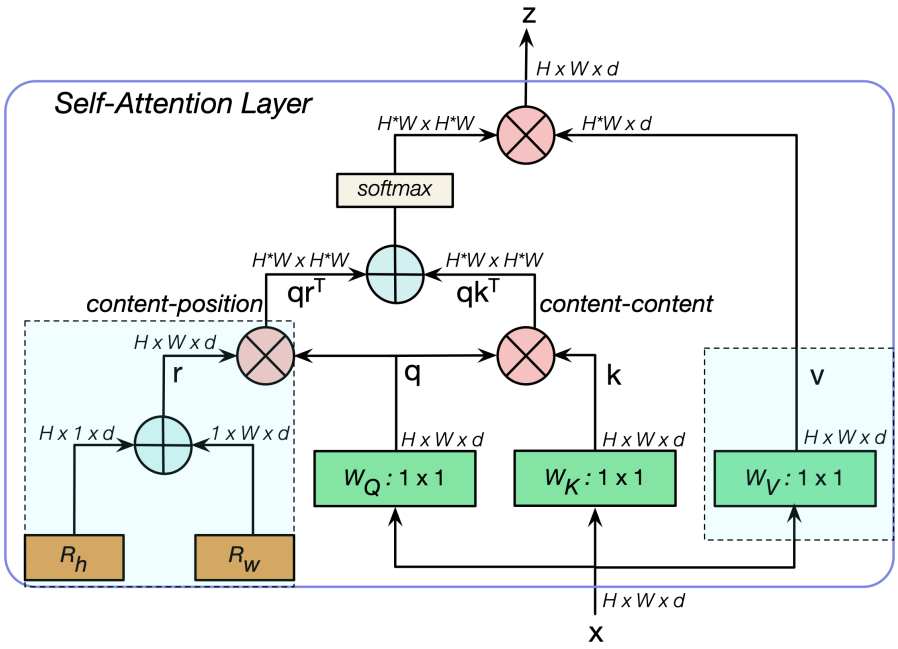


Figura 3.10: MSA nel BoTNet50 [1]

Per quanto riguarda la codifica posizionale, viene impiegato un bias posizionale relativo [22, 1] in modalità simile a quella descritta in 3.2.5, tenendo conto del fatto che in questo caso la MSA è globale e non limitata da finestre. Il bias è calcolato in base alla distanza relativa tra ciascuna coppia di posizioni nelle feature map convoluzionali. È stato osservato che le codifica posizionale basata sulla distanza relativa sono più adatte per i task di visione [1, 24, 25, 26]. Questo può essere

stage	output	ResNet-50	BoTNet-50
c1	112×112	$7 \times 7, 64$, stride 2 3×3 max pool, stride 2	$7 \times 7, 64$, stride 2 3×3 max pool, stride 2
c2	56×56	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
c3	28×28	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
c4	14×14	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$
c5	7×7	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ \mathbf{MSA}, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
# params.		25.5×10^6	20.8×10^6

Tabella 3.1: Tabella comparativa tra ResNet-50 e BoTNet-50 con input di dimensioni 224×224.

attribuito al fatto che l'attenzione non solo considera le informazioni sul contenuto, ma anche le distanze relative tra le feature in diverse posizioni, permettendo così di associare efficacemente le informazioni agli oggetti con consapevolezza sulla loro posizione [1].

3.4 Procedura sperimentale

In questo capitolo vengono descritti in dettaglio l'ambiente e le metodologie adottate per l'addestramento dei modelli selezionati. L'obiettivo è di delineare le scelte tecniche relative all'addestramento. Verranno analizzate le decisioni prese in termini di selezione della funzione di perdita, configurazione dell'ottimizzatore, le tecniche di regolarizzazione e metriche di valutazione, fornendo una completa panoramica dell'infrastruttura e delle strategie di addestramento impiegate.

3.4.1 Funzione di perdita

La tesi esamina un problema di classificazione che coinvolge quattro classi, corrispondenti ai diversi livelli di gravità, come delineato nella sezione 3.1.3. Queste classi rappresentano i vari score di gravità associati ai pazienti valutati tramite immagini ecografiche. Dato l'evidente problema di sbilanciamento delle classi nel dataset è stata selezionata la *Weighted Categorical Cross Entropy* (WCCE) come funzione di loss. La WCCE è una variante della tradizionale Categorical Cross Entropy (CCE)

che introduce un sistema di pesi per gestire i dataset con distribuzioni di classe sbilanciate. L'utilizzo dei pesi è necessario per mitigare gli effetti dello sbilanciamento delle classi nel dataset, visto in Figura 3.3.

$$WCCE = -\sum_{i=1}^{N}\sum_{c=1}^{C} w_c \cdot y_{ic} \cdot \log(\hat{y}_{ic})$$

dove:

- N rappresenta il numero di campioni.
- C è il numero di classi.
- w_c è il peso assegnato alla classe c.
- y_{ic} è un indicatore binario che denota se la classe c è corretta per l'i-esimo campione.
- $\hat{y}_{ic}$ è la probabilità prevista per la classe c dell'i-esimo campione.

Per convertire i logit (output grezzi del modello) in probabilità viene utilizzata la softmax. Per un vettore di logit $\mathbf{z} = [z_1, z_2, \ldots, z_C]$, dove C rappresenta il numero di classi, la softmax è definita come:

$$\text{softmax}(\mathbf{z})_c = \frac{e^{z_c}}{\sum_{j=1}^{C} e^{z_j}} \quad \text{per } c = 1, 2, \ldots, C$$

La funzione garantisce che la somma delle probabilità predette per tutte le classi sia uguale a 1 e che ciascuna probabilità sia compresa tra 0 e 1.

I pesi w_c sono generalmente calcolati in base alla frequenza inversa delle classi nel dataset, favorendo le classi meno rappresentate. L'errore relativo a ciascuna classe viene enfatizzato essendo moltiplicato per il suo peso associato. Le classi meno frequenti ricevono un peso maggiore, quindi gli errori su queste classi hanno un impatto maggiore sulla modifica dei pesi della rete.

3.4.2 Ottimizzatore SGD

Lo Stochastic Gradient Descent (SGD) è un metodo di ottimizzazione ampiamente utilizzato nell'ambito dell'apprendimento automatico. A differenza del tradizionale Gradient Descent, che calcola il gradiente della funzione di perdita sull'intero dataset, SGD aggiorna i parametri del modello utilizzando un sottoinsieme (batch) dei dati, rendendolo più efficiente per dataset di grandi dimensioni.

Il processo di aggiornamento dei parametri attraverso SGD è definito come segue:

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t; X_{\text{batch}}, Y_{\text{batch}})$$

dove θ rappresenta i parametri del modello, η è il tasso di apprendimento (o learning rate), e ∇L denota il gradiente della funzione di perdita L (nel nostro caso la WCCE), calcolato sul batch di dati $(X_{\text{batch}}, Y_{\text{batch}})$.

Momentum in SGD

Il concetto di momentum è stato introdotto per mitigare alcuni dei limiti dell'SGD, come la lentezza nella convergenza e la vulnerabilità ai minimi locali [27]. Il momentum accelera l'SGD nella direzione del gradiente decrescente e smorza le oscillazioni, permettendo un addestramento più rapido e stabile. L'aggiornamento dei parametri con momentum è dato da:

$$\mathbf{v}_{t+1} = \mu_t \mathbf{v}_t - \eta \nabla L(\theta_t)$$

$$\theta_{t+1} = \theta_t + \mathbf{v}_{t+1}$$

dove $\mathbf{v}_t$ è il vettore di momentum e μ è il coefficiente di momentum. Tipicamente, μ è impostato a un valore vicino a 1, come 0.9, per garantire che il momentum abbia un impatto significativo sull'aggiornamento dei parametri. L'introduzione del momentum in SGD aiuta a superare i plateau della loss e a evitare i minimi locali, contribuendo a una convergenza più veloce.

Anche se più lento nella convergenza, l'SGD è stato scelto come ottimizzatore per l'addestramento dei modelli da zero in quanto tende a essere più stabile e porta generalmente a una generalizzazione migliore. Ciò assume un certo grado di rilevanza vista la scarsità e la complessità del dataset.

3.4.3 Ottimizzatore AdamW

Per la fase di transfer learning, si è preferito l'ottimizzatore AdamW in quanto utilizzato per l'allenamento delle architetture sul dataset ImageNet. AdamW è una variante dell'ottimizzatore Adam che incorpora la tecnica di weight decay direttamente nel processo di ottimizzazione. Mentre lo Stochastic Gradient Descent (SGD) aggiorna tutti i parametri con lo stesso tasso di apprendimento, Adam è noto per la sua efficienza nell'aggiustare il learning rate per ciascun parametro basandosi sui momenti del primo e secondo ordine delle stime dei gradienti. AdamW modifica l'approccio standard del weight decay, rendendolo più compatibile con la metodologia di aggiornamento dei parametri di Adam. Questo rende AdamW particolarmente efficace per problemi con spazi di parametri grandi e complessi, dove la sintonizzazione del learning rate per parametro può avere un impatto significativo sulla convergenza e sulle prestazioni del modello.

3.4.4 Scheduler

Gli Scheduler nell'Apprendimento Automatico

Nell'apprendimento automatico, uno scheduler di apprendimento è un metodo per regolare il learning rate (tasso di apprendimento) durante l'addestramento di un modello. Lo scheduler modifica il learning rate in base a una politica predeterminata, influenzando la velocità e la qualità della convergenza del modello. L'uso di uno scheduler è fondamentale per bilanciare la velocità di apprendimento e la capacità del modello di trovare un minimo globale ottimo nella funzione di loss [28].

Cosine Annealing with Warm Restarts

Il Cosine Annealing with Warm Restarts è un scheduler ciclico introdotto in [29] che adatta il tasso di apprendimento seguendo una funzione coseno che diminuisce progressivamente nel tempo, migliorando la convergenza del modello. Inoltre, periodicamente vengono eseguiti dei riavvii, o warm restarts, che ripristinano il tasso di apprendimento al valore iniziale (Figura 3.11). I riavvii permettono di uscire dai minimi locali sub ottimali portando a "un'esplorazione" dello spazio dei parametri e della *loss landscape* in modo simile a come mostrato nella Figura 3.12 [30]. Il tasso di apprendimento alla i-esima iterazione è definito come segue [29]:

$$\eta_i = \eta_{\min}^i + \frac{1}{2}(\eta_{\max}^i - \eta_{\min}^i)\left(1 + \cos\left(\frac{T_{\text{cur}}}{T_i}\pi\right)\right)$$

dove:

- η_i è il tasso di apprendimento alla i-esima iterazione.
- $\eta_{\min}^i$ e $\eta_{\max}^i$ sono rispettivamente i limiti inferiore e superiore per il tasso di apprendimento alla i-esima iterazione.
- T_{cur} è il numero di iterazioni dall'ultimo restart.
- T_i è il numero totale di iterazioni tra due restart.

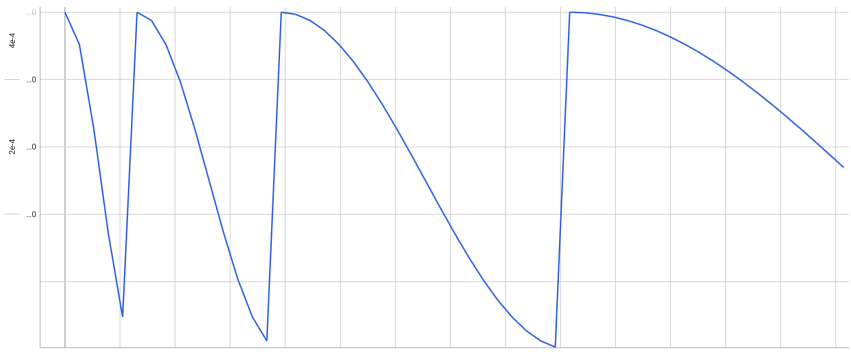


Figura 3.11: Andamento del learning rate durante l'addestramento con lo scheduler Cosine Annealing with Warm Restarts

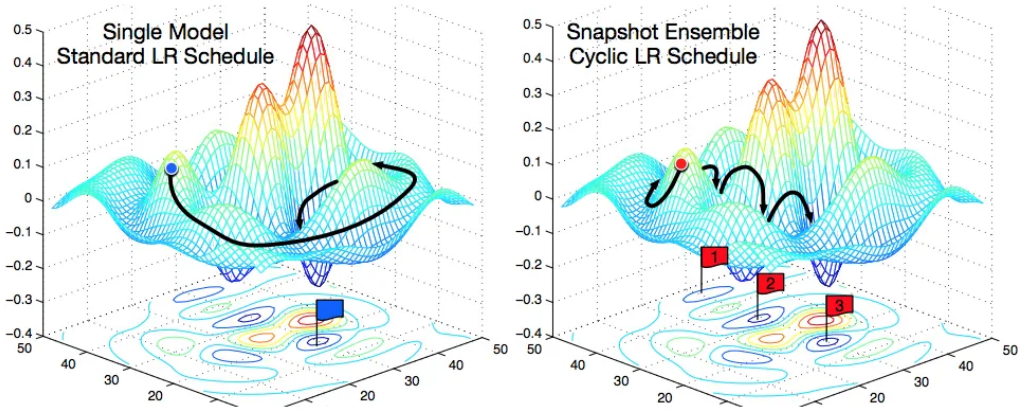


Figura 3.12: Confronto tra gli effetti di uno scheduler ciclico (a destra) e non, sull'ottimizzazione della loss per il raggiungimento del minimo ottimale

3.4.5 Regolarizzazione

La regolarizzazione comprende un insieme di approcci volti a migliorare la capacità generalizzativa prevenendo l'overfitting, una condizione in cui il modello si adatta eccessivamente ai dati di addestramento. La regolarizzazione rappresenta perciò, un aspetto cruciale nell'addestramento di modelli deep learning. Viste anche le caratteristiche del dataset ICLUS, in cui i dati sono limitati e con un certo grado di complessità, l'overfitting è particolarmente problematico in quanto piuttosto che le sottostanti distribuzioni o pattern, il modello può facilmente apprendere il rumore presente nei dati. La regolarizzazione ha quindi come scopo quello di spingere il modello a imparare solo le caratteristiche più rilevanti e robuste portando a una migliore generalizzazione e prestazioni più stabili su dati non visti.

Nel contesto di questo elaborato, sono stati adottati come tecniche di regolarizzazione il *weight decay*, il *dropout* la *data augmentation* e *l'early stop* (affrontata in 3.4.9).

Weight Decay

Il *weight decay* è una forma di regolarizzazione L2 che aggiunge un termine di penalità proporzionale al quadrato dei parametri del modello alla funzione di perdita. Questo approccio tende a ridurre la magnitudine dei parametri evitando anche l'esplosione dei gradienti e portando a una migliore generalizzazione. Il termine di penalità è espresso come:

$$L_{\text{weight decay}} = \lambda \sum_i \theta_i^2$$

dove λ è un iperparametro di regolarizzazione che controlla l'intensità del weight decay e θ_i rappresenta i parametri del modello. Il valore di λ deve essere tale da

portare al giusto compromesso tra generalizzazione e complessità senza causare overfitting o underfitting.

Dropout

Il *dropout* (Figura 3.13) è una tecnica di regolarizzazione specifica per il deep learning che aiuta a prevenire la co-adattabilità dei neuroni nella rete. Durante la fase di addestramento, il dropout seleziona casualmente un certo numero di unità (neuroni) in uno strato della rete e riduce temporaneamente i loro valori a zero. Questo processo, che può essere visto come la creazione di una versione "snellita" del modello a ogni iterazione, impedisce ai neuroni di sviluppare una dipendenza eccessiva l'uno dall'altro, spingendo invece la rete a imparare caratteristiche più robuste e ridondanti. Il tasso di dropout, che indica la percentuale di neuroni da disattivare, è un parametro chiave che può essere ottimizzato per bilanciare l'effetto della regolarizzazione.

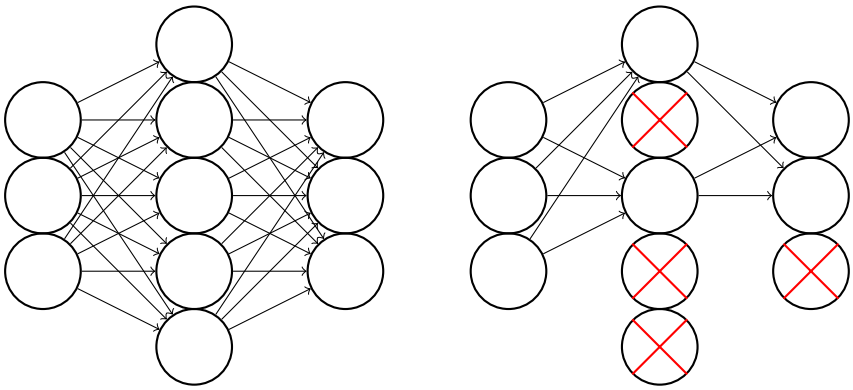


Figura 3.13: Rete neurale con e senza dropout. I nodi contrassegnati con una croce rossa rappresentano i neuroni disattivati.

Data Augmentation

La *data augmentation* è una strategia per aumentare la varietà e la quantità di dati utilizzati in fase di addestramento. È un approccio molto utilizzato in contesti dove la raccolta di dataset grandi o variegati è difficile o costosa. Nello specifico, la data augmentation comporta l'alterazione casuale dei dati di addestramento esistenti per creare nuovi campioni "artificiali". Nel caso delle immagini può includere varie trasformazioni come rotazioni, traslazioni, riflessioni, rumore aggiuntivo, variazioni di luminosità e contrasto, tra le altre. Nel contesto del deep learning per il riconoscimento d'immagini, la data augmentation è particolarmente preziosa, poiché le reti deep, tra cui le cnn, possono essere molto sensibili alle variazioni nelle posizioni e nelle orientazioni degli oggetti nelle immagini. L'implementazione di tecniche di data augmentation aiuta a costruire modelli che sono meno suscettibili a tali variazioni, migliorando la loro robustezza e versatilità.

Nel contesto specifico di questa tesi, la data augmentation sul dataset ICLUS è stata implementata in modalità *online*. Le trasformazioni (Tabella 3.2) sono sia geometriche che a livello dei pixel e sono clinicamente significative e plausibili nel contesto del dataset ICLUS [9]. L'applicazione avviene in maniera casuale e in tempo reale sulle immagini a ogni iterazione. Viene introdotto così un certo grado di varietà garantendo che il modello non veda mai (o molto raramente) due volte lo stesso esatto campione durante l'allenamento, migliorando così la sua capacità di generalizzare a nuovi dati e riducendo il rischio di overfitting.

Trasformazione	**Valore**
Rotazione	$\pm$ 20°
Fattore di scala	[1, 1.5]
Traslazione	[-0.15, 0.15]
Flip Orizzontale	$p = 50\%$
Luminosità	[-0.3, 0.3]
Contrasto	[-0.3, 0.3]
Tonalità	[-0.3, 0.3]

Tabella 3.2: Trasformazioni con i rispettivi valori

3.4.6 Metriche

In questa sezione vengono descritte le metriche utilizzate per valutare le prestazioni dei modelli, considerando in particolare il problema dello squilibrio delle classi nel dataset. Le metriche chiave includono F1-score, accuratezza e sensibilità.

F1-score

Il F1-score è una misura che combina precisione e sensibilità. In contesti in cui è presente un sbilanciamento delle classi è da preferire rispetto all'accuratezza poiché considera sia i falsi positivi sia i falsi negativi. È definita come:

$$F1 = 2 \times \frac{\text{Precisione} \times \text{Sensibilità}}{\text{Precisione} + \text{Sensibilità}}$$

Precisione

La precisione è la frazione dei veri positivi rispetto al totale dei casi classificati come positivi. Valuta quanti dei casi classificati come positivi dal sono realmente positivi. È particolarmente importante in situazioni dove il costo dei falsi positivi (FP) è elevato. È definita come:

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

Sensibilità

La sensibilità o recupero (recall), rappresenta il tasso dei veri positivi, ovvero misura la proporzione dei positivi reali che vengono identificati correttamente. È cruciale in applicazioni mediche dove il mancato riconoscimento di un caso positivo può essere critico.

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

Accuratezza

L'accuratezza è la frazione di previsioni corrette tra il totale delle previsioni. Tuttavia, in contesti con classi sbilanciate, questa metrica può essere fuorviante, poiché favorisce la classe maggioritaria.

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

3.4.7 ROC, AUROC e matrici di confusione

Le performance dei modelli sono valutate utilizzando anche le curve ROC (Receiver Operating Characteristic) e le matrici di confusione. La ROC è uno strumento grafico che offre una rappresentazione visiva della capacità di un classificatore di distinguere tra le classi, mostrando la relazione tra la sensibilità, o Tasso di Veri Positivi (TPR), e il Tasso di Falsi Positivi (FPR). In aggiunta viene indicato anche l'indice AUC (Area Under Curve), che misura l'area sottesa da questa curva. Nel caso specifico della classificazione multiclasse, si adotta l'approccio One vs the Rest (OvR), analizzando separatamente la capacità del modello di identificare ciascuna classe contro tutte le altre. La metrica AUROC riportata nelle tabelle dei risultati viene invece calcolata come la media pesata (visto lo sbilanciamento delle classi) delle AUC per ciascuna classe.

Le matrici di confusione offrono invece una visione dettagliata delle prestazioni di classificazione permettendo di rilevare il numero di predizioni corrette ed errate per ciascuna classe. Perciò, sono un ottimo modo per identificare tipologie specifiche di errori, quali la tendenza a confondere particolari classi, fornendo così intuizioni preziose sul comportamento dei modelli.

3.4.8 GradCAM - Valutazione qualitativa

Le Grad-CAM sono un metodo interpretativo che permette di visualizzare quali regioni dell'immagine sono considerate importanti dal modello per prendere una decisione di classificazione. Il processo di base per l'ottenimento delle Grad-CAM è il seguente:

1. Si calcolano i gradienti del punteggio di ciascuna classe rispetto alle feature map dell'ultimo strato. Questi gradienti sono poi mediati globalmente per ottenere i pesi.

2. Si esegue una combinazione pesata delle feature map con i pesi calcolati, seguita da una operazione di ReLU. Questo produce una heatmap per ciascuna classe, che evidenzia le regioni più importanti per quella classe.

Nel caso del Swin Transformer è necessario adattare il processo in modo che si considerino i gradienti rispetto alle attention map generate dal MSA, calcolando poi una media ponderata basata su queste per produrre le heatmaps.

L'applicazione delle Grad-CAM in questo studio tenta di fornire una visualizzazione interpretabile di quali parti dell'immagine influenzano maggiormente le decisioni del modello (Figura 3.14). Questo non solo aiuta a comprendere meglio il modello stesso, ma fornisce anche una valida verifica per assicurarsi che il modello si focalizzi sulle regioni giuste delle immagini per la classificazione, un aspetto particolarmente critico in applicazioni mediche e diagnostici dove la localizzazione delle aree d'interesse è essenziale.

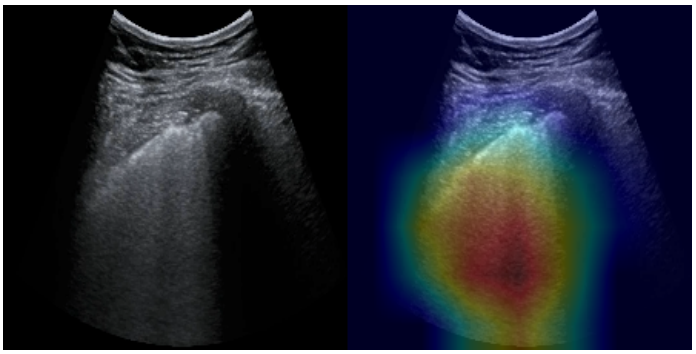


Figura 3.14: Esempio di GradCAM

3.4.9 Esperimenti

Questa sezione offre una breve panoramica della procedura di svolgimento degli esperimenti. Nella loro conduzione, si adotta un approccio metodico per ottimizzare la configurazione di allenamento di ciascun modello in modo da massimizzarne le prestazioni e ridurre il più possibile l'overfitting.

Tutti gli esperimenti sono stati eseguiti su una configurazione hardware dotata di due CPU Intel Xeon Silver 4214 e una GPU RTX 2080 Ti con 11 GB di memoria video GDDR6, 544 Tensor Core, 4352 CUDA core.

Split del dataset I dati vengono inizialmente suddivisi in set di allenamento (train), validazione (val) e test, seguendo proporzioni 60/20/20. Inoltre ci si assicura che i set abbiano una distribuzione delle classi simile a quella del dataset completo. Questo approccio mira a riflettere la distribuzione reale delle classi nel mondo reale. La

rappresentazione proporzionata delle classi in ogni set assicura che il modello venga testato e validato in condizioni simili a quelle in cui sarà utilizzato.

Dimensione del Batch Invece di utilizzare l'intero trainset o un singolo esempio per calcolare il gradiente, il set viene suddiviso in piccoli gruppi chiamati "mini-batch". Il ciclo di addestramento viene eseguito iterando su questi mini-batch. La dimensione di quest'ultimi è un iperparametro che può variare a seconda delle specifiche del problema e delle limitazioni hardware Questo permette di avere un equilibrio tra l'efficienza della memoria e la velocità di calcolo, riducendo al tempo stesso la varianza degli aggiornamenti dei pesi. In questo caso si è optato per l'utilizzo una dimensione di batch fissa a 64 campioni. Questa scelta è dettata dalla volontà di sfruttare al massimo la capacità della memoria video disponibile.

Early Stop Per prevenire l'overfitting e migliorare la generalizzazione, si adotta una tecnica di early stopping per interrompere l'addestramento quando la performance in termini di f1-score sul set di validazione smette di migliorare, un segnale che il modello ha iniziato a specializzarsi eccessivamente sui dati di addestramento. L'early stopping è fondamentale per garantire che il modello mantenga un livello adeguato di generalizzazione.

Ottimizzazione con Grid Search e Testing Nella pipeline di addestramento (Figura 3.15), si impiega una strategia di ottimizzazione basata sull'uso del *Grid Search* per la ricerca dei valori ottimali degli iperparametri critici quali il learning rate, il dropout e il weight decay. Per ciascun modelli sono messe a disposizioni 20 esecuzioni totali, con un numero massimo di 30 epoche, per provare le varie configurazioni. Al termine del processo di ottimizzazione, si seleziona la configurazione del ciclo di addestramento con le migliori prestazioni sul set di validazione per iniziare un ciclo di addestramento più esteso fino al raggiungimento del numero massimo di epoche oppure fino all'interruzione da parte dell'early stopper. Durante questa fase vengono salvati i tre migliori checkpoint del modello in base al f1-score sul set di validazione. Una volta completato il processo si utilizza uno dei checkpoint per procedere con il test finale. Da sottolineare che la selezione mediante Grid Search viene applicata per il weight decay e il dropout, mentre il tasso di apprendimento viene campionato da una distribuzione log-uniforme definita da un intervallo di valori possibili (ad esempio $[10^{-4}, 0.01]$). Questo implica che ogni ordine di grandezza ha la stessa probabilità di essere scelto e ciò è utile per esplorare valori di learning rate sia molto piccoli che molto grandi in modo equilibrato. Finita la fase di test, i risultati ottenuti vengono utilizzati per un confronto tra ResNet50, BoTNet50 e Swin.

Questo processo garantisce che i modelli operino con la configurazione ottimale in modo da ottener il miglior equilibrio tra apprendimento e capacità di generalizzazione tramite, eseguendo controllo rigoroso del rischio di overfitting e garantendo così la validità e l'affidabilità dei risultati ottenuti.

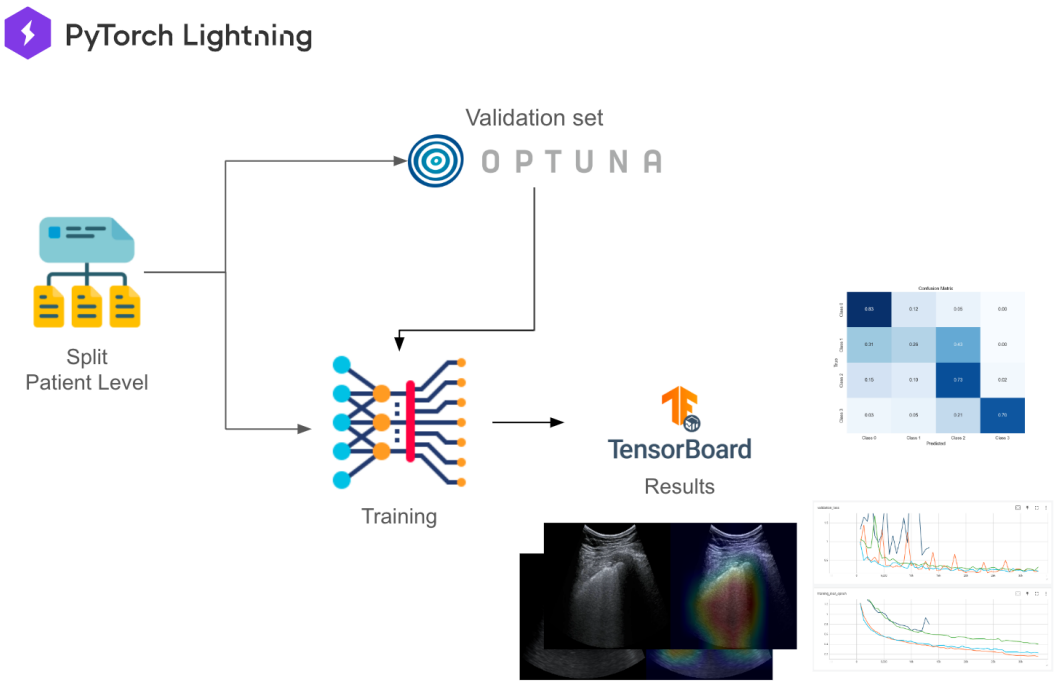


Figura 3.15: Panoramica delle framework sviluppato

Capitolo 4

Risultati e Discussioni

In questo capitolo, sono presentati i risultati degli esperimenti condotti per valutare l'applicabilità e l'efficacia dei modelli basati su Transformer nell'analisi di immagini ecografiche polmonari, confrontandoli con CNN di riferimento, ResNet50. L'indagine si è concentrata su due approcci principali: l'impiego di un'architettura basata esclusivamente su meccanismi di Multi-Head Self Attention (MSA) senza convoluzioni (Swin) e l'integrazione dell'MSA all'interno di una backbone convoluzionale per creare un modello ibrido (BoTNet50).

L'obiettivo principale è di determinare se questi approcci hanno le potenzialità di diventare backbone per l'ambito specifico dell'imaging polmonare, considerando la capacità di classificare correttamente i casi più critici e la generalizzabilità dei modelli. Un focus particolare è stato posto sulla limitata disponibilità di dati nel dataset, valutando l'impiego del transfer learning come strumento per ottimizzare l'addestramento dei modelli basati su pura attenzione e migliorare le loro performance sul dataset ICLUS. Inoltre, è stata esaminata l'interpretabilità dei modelli e la loro capacità di fornire risultati trasparenti e comprensibili ai professionisti medici, un aspetto fondamentale per l'accettazione clinica.

I tempi di addestramento si aggiravano mediamente tra i 4 - 6 minuti a epoca per tutti i modelli sotto analisi. I risultati ottenuti mettono in luce differenze significative nelle prestazioni tra i tre modelli. Nonostante i tre modelli presentino un numero paragonabile di parametri addestrabili, ciò non si traduce in prestazioni uniformi, evidenziando come l'architettura specifica di ciascun modello giochi un ruolo critico nell'apprendimento delle caratteristiche dalle immagini.

È importante ribadire che nello stato dell'arte, per il dataset ICLUS, non esistono modelli MSA che possano fungere da baseline, rendendo questo studio unico nell'esplorare l'efficacia dei transformer per questo dataset. Inoltre, non è possibile stabilire un confronto diretto con i lavori precedenti di Roy et al. [8] e Frank et al. [9], a causa dell'indisponibilità degli holdout utilizzati in quei studi. Pertanto, piuttosto che mirare a superare le prestazioni dei modelli dello stato dell'arte, lo studio cerca di confrontare l'efficacia dei meccanismi di Multi-Head Self Attention (MSA) rispetto alle architetture CNN tradizionali dei quali la ResNet50 è stata presa come architettura di riferimento.

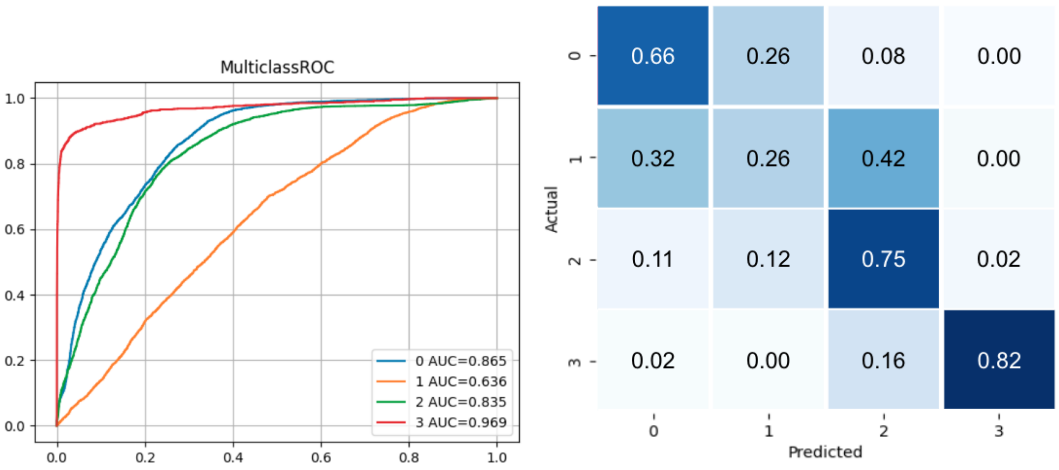
Nei paragrafi seguenti, vengono discussi dettagliatamente i risultati ottenuti dai modelli impiegati, mettendo in luce le loro prestazioni, i vantaggi e le limitazioni nell'ambito dell'analisi ecografica polmonare.

4.1 Confronto backbone allenate da zero

L'analisi comparativa è stata effettuata utilizzando una configurazione sperimentale in cui i modelli sono stati addestrati da zero, *from scratch*, per valutare le capacità intrinseche di ResNet50, BoTNet50 e Swin Tiny nell'elaborazione di immagini ecografiche polmonari. Nella tabella 4.1 vengono riportati gli iperparametri e i risultati ottenuti delle configurazioni emerse vincenti dal processo di ottimizzazione con Grid Search.

	ResNet50	**BoTNet50**	**Swin Tiny**
LR iniziale	0.0007	0.0039	0.0001
Dropout	0.1	0.2	0.3
Weight Decay	0.01	0.01	0.01
Numero di parametri	25.6 M	20.8 M	27.5 M
AUROC	0.7884	**0.8057**	0.7055
Accuracy	**0.6013**	0.5974	0.4649
F1-Score	0.5896	**0.6025**	0.4682

Tabella 4.1: Confronto delle prestazioni dei modelli senza transfer learning.



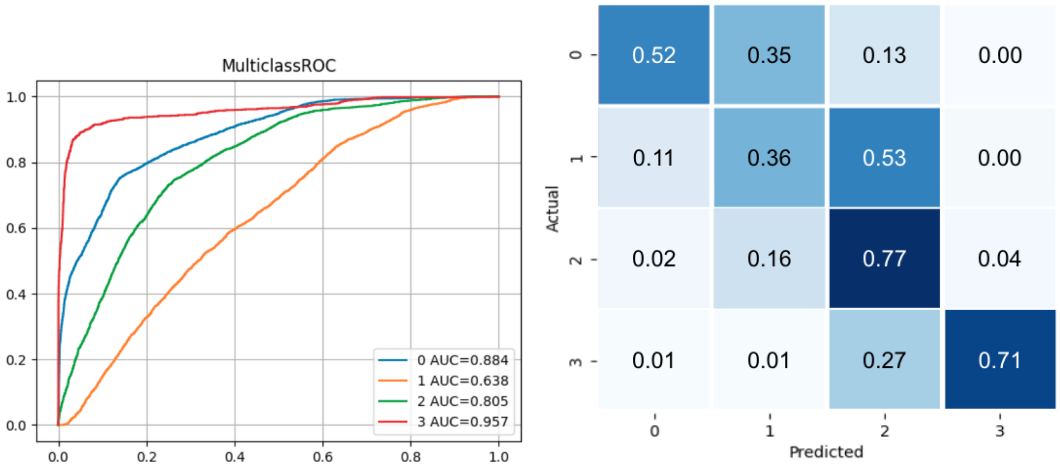
(a) ROC (b) Matrice di confusione

Figura 4.1: Curve ROC e matrice di confusione di ResNet50

ResNet50 ha dimostrato una solida prestazione generale con un F1-score di 0.5896. Questi risultati sono in linea con quelli riscontrati in [8] e [9]. Osservando la Figura 4.1

la matrice di confusione mostra una predominanza nella classificazione corretta delle istanze della classe 3, con un valore di 0.82, indicando un'alta precisione per questo score. Tuttavia, si riscontra una difficoltà significativa nel identificare correttamente i frame di score 1 da quelli di classe 0 e 2. Ciò è riscontrabile anche osservando le curve ROC (Receiver operating characteristic) dal quale si può osservare che l'AUC per la classe 1 è di gran lunga inferiore alle altre classi.

Passando a **BoTNet50**, che integra un meccanismo di self-attention negli strati più profondi della ResNet, ha superato quest'ultima in termini di F1-score con un valore di 0.6025 e ha mostrato un miglioramento nell'AUROC raggiungendo 0.8057, il che suggerirebbe una maggiore capacità di discriminazione. Il modello mostra un miglioramento del 10% per la classe 1 e del 2% per la classe 2. È particolarmente rilevante notare che il modello manifesta una propensione a classificare erroneamente i campioni sani come appartenenti alle classi 1 o 2, che corrispondono a condizioni di maggiore gravità. Questa inclinazione può portare a un aumento dei falsi positivi che nel contesto clinico della malattia COVID-19 comporterebbe un maggior grado di cautela e ulteriori accertamenti. Inoltre vi è un peggioramento significativo del 11% per quanto riguarda la classe 3. È plausibile pensare che il fatto che la classe 3 sia quella minoritaria, abbia influito maggiormente sul modello rispetto alla ResNet per la presenza della MSA negli ultimi strati dell'architettura . In generale, l'introduzione del MSA ha portato un miglioramento nella classificazione delle classi intermedie ma dei risultati inferiori per le classi estreme 3 e 0, oltre alla tendenza, per quest'ultima, a commettere errori di tipo 1 che a seconda del contesto clinico può essere un punto a favore o meno.



(a) ROC (b) Matrice di confusione

Figura 4.2: Curve ROC e matrice di confusione di BoTNet50

La tabella 4.2 mostra le performance delle migliori tre configurazioni, dopo la vincente i cui risultati sono stati riporati sopra, raggiunte da BoTNet50 e ResNet50 durante la fase di ottimizzazione Grid Search sullo stesso set di validazione. Dall'analisi della tabella, si può osservare che BoTNet50 sembra raggiungere performance

più consistenti rispetto a ResNet50. In particolare, BoTNet50 mantiene un livello di performance relativamente stabile attraverso diverse configurazioni, suggerendo una migliore robustezza del modello alle variazioni dei parametri di addestramento. Questo potrebbe essere attribuito all'integrazione del meccanismo di self-attention, suggerendo anche che tale integrazione possa conferire una robustezza superiore contro le variazioni degli iperparametri di addestramento. In confronto, ResNet50 ha mostrato una maggiore sensibilità alle modifiche nei parametri, come evidenziato dal calo di performance tra la prima e la terza configurazione.

Modello	**Configurazione 1**	**Configurazione 2**	**Configurazione 3**
BoTNet50	**0.6454**	**0.6253**	**0.6210**
ResNet50	0.6245	0.6095	0.5843

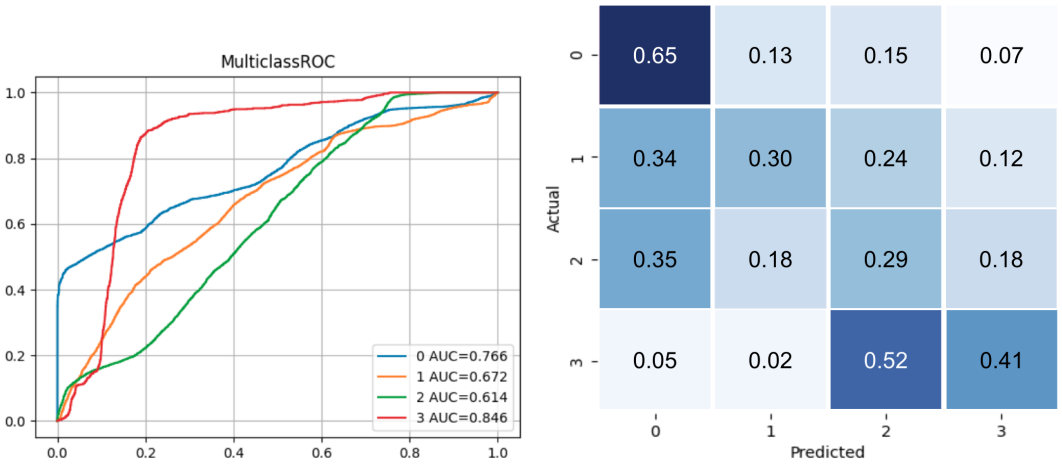
Tabella 4.2: Performance sul validation set di BoTNet50 e ResNet50 nelle migliori tre configurazioni dopo la vincente.

4.2 Considerazioni sulle performance di Swin-T e sui bias induttivi

Lo Swin Tiny, che si affida esclusivamente al meccanismo di self-attention senza incorporare strati convoluzionali, non ha raggiunto livelli di performance comparabili con i modelli che includono convoluzioni. Anche dando una rapida occhiata alle curve ROC e alla matrice di confusione nella Figura 4.3 si può vedere la notevole difficoltà del modello, sopratutto per quanto riguarda le classi intermedie.

Una possibile causa di questa discrepanza potrebbe essere attribuita alla mancanza nei Transformer di alcuni bias indotti dalle convoluzioni che nelle CNN aiutano a catturare le relazioni spaziali locali e le gerarchie di pattern nelle immagini. Le convoluzioni, infatti, introducono dei bias architetturali tipici, come l'equivarianza alle traslazioni, che predispongono il modello a favorire l'apprendimento di una determinata tipologia di feature locali [16]. Questo può essere particolarmente vantaggioso in scenari con dati limitati, poiché il modello è "accompagnato" verso una classe specifica di caratteristiche che si è storicamente dimostrata efficace per compiti visivi portando a un apprendimento più veloce e una miglior capacità generalizzativa. Questi stessi bias, tipici delle convoluzioni, però porterebbero ad un apprendimento meno flessibile e più limitato quando la quantità di dati è maggiore [31]. Nel Swin, a differenza del ViT, alcuni di questi bias vengono in parte recuperati grazie all'approccio gerarchico nel applicare il MSA [19] ma i risultati suggeriscono che non sono sufficienti in un contesto con un dataset clinico e risorse computazionali limitate come quello presentato. Da qui la riconferma di ciò che si è riscontrato in letteratura [16, 14], ovvero che le architetture vision transformer basate unicamente su self-attention, sebbene possano teoricamente catturare dipendenze a lungo raggio

all'interno dei dati e portare performance migliori proprio per la minor presenza di questi bias rispetto alle CNN, per farlo potrebbero richiedere un volume di dati maggiore e tempi più estesi per apprendere efficacemente tali relazioni complesse da zero.correlazioni da zero.



(a) ROC (b) Matrice di confusione

Figura 4.3: Curve ROC e matrice di confusione di Swin Tiny

Per superare questa limitazione, sono stati esplorati diversi metodi per incorporare dei bias nei modelli pure attention, in modo tale però da preservare la loro intrinseca flessibilità data dal meccanismo di self-attention. L'efficacia della convoluzione si basa sul fatto che i pixel vicini nelle immagini naturali sono altamente correlati, ma possono esistere altri contenuti altamente correlati al di fuori del campo recettivo locale di un filtro convoluzionale che vengono ignorati [32]. Lee et al. [31] hanno proposto un metodo di riparametrizzazione progressiva per modulare il bias indotto tra la convoluzione e self-attention, permettendo ai modelli di adattarsi meglio alla grandezza del dataset. Zhang et al. [33] hanno osservato che i ViT generalizzano meglio delle CNN sotto variazioni di distribuzione, grazie a bias più inclini verso forme e strutture piuttosto che background e texture. Zhou et al. [32] hanno introdotto l'uso di Spatial Prior-enhanced Self-Attention (SP-SA) consentendo ai ViT di apprendere autonomamente durante l'addestramento una varietà di bias che enfatizzano certe relazioni spaziali. Questi approcci mostrano come l'introduzione di bias possano aiutare i modelli non dotati di convoluzioni a migliorare la loro efficienza e potenziale di generalizzazione su dataset di scala inferiore.

4.3 Transfer Learning

L'applicazione dei transformer visuali nel settore medico, in particolare nell'analisi delle immagini ecografiche, rappresenta un campo di ricerca emergente con un grande potenziale. Attualmente, l'aggiunta di bias induttivi in questi modelli è una direzione di studio ancora poco esplorata nel contesto medico. L'introduzione di tali bias

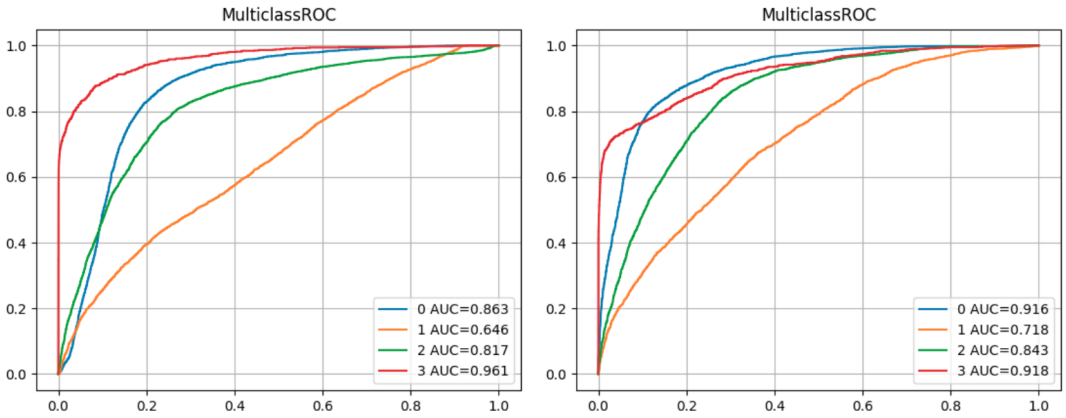
potrebbe essere particolarmente vantaggiosa ma richiede ulteriori studi supportati anche dagli esperti di dominio. Nel frattempo, un approccio ampiamente adottato in ambito medico è il **transfer learning**, tecnica che consiste nell'utilizzare modelli pre-addestrati su dataset vasti e generali per poi adattarli a dataset più specifici e circoscritti come quelli clinici. Attraverso questo processo il Swin può beneficiare dell'apprendimento pregresso da una vasta gamma di dati visivi e, successivamente, essere raffinato per riconoscere le caratteristiche peculiari delle immagini ICLUS. Questo è il metodo adottato per mitigare il problema dato dalla scarsità di dati e risorse computazionali.

La tabella sottostante riflette le prestazioni di ResNet50 e Swin Tiny pre-allenate su ImageNet-1k(1000 classi su 1,281,167 immagini). Sfortunatamente il BoTNet50 non è incluso nel confronto per l'indisponibilità di un modello pre-allenato. Come già accennato in 3.4.3, per questo confronto è stato preferito AdamW al SGD in quanto ottimizzatore utilizzato per l'allenamento delle architetture sul dataset ImageNet. A entrambi i modelli è stata sostituita la "testa" di classificazione con una nuova adatta per il problema di classificazione a quattro classi.

	ResNet50	**Swin Tiny**
LR iniziale	0.001	0.002
Dropout	0.1	0.1
Weight Decay	0.01	0.001
AUROC	0.8017	**0.8311**
Accuracy	0.6388	**0.6594**
F1-Score	0.6280	**0.6513**

Tabella 4.3: Risultati del transfer learning dal dataset ImageNet-1k.

Come mostrato, entrambi i modelli hanno beneficiato significativamente del transfer learning, ottenendo un miglioramento sia in termini di accuratezza che di F1-Score. Nel caso di ResNet50 si attestano rispettivamente al 63.88% e al 62.80%. Permangono le difficoltà nel distinguere la classe 1 dalle altre, come si può osservare nella Figura 4.4. Ma è con il Swin Tiny che il processo ha avuto maggiori effetti. Nonostante la sua natura esente da convoluzioni, Swin Tiny ha superato ResNet50 su tutte le metriche. AUROC con un punteggio di 0.8311, ottenendo AUC migliori rispetto a ResNet50 per tutte le classi ad eccezione della 3, in modo similare a ciò che si è visto con BoTNet50 in 4.1, suggerendo nonostante ciò una migliore capacità nella differenziazione delle classi. In termini di accuratezza e F1-Score, Swin Tiny ha ottenuto rispettivamente il 65.94% e il 65.13%, superando ResNet50 e sottolineando l'efficacia del transfer learning anche per i modelli basati su self-attention.



(a) ResNet50 (b) Swin Tiny

Figura 4.4: Curve ROC di ResNet50 e Swin Tiny pre-allenati su ImageNet-1k

I risultati ottenuti attraverso il transfer learning sottolineano l'importanza di questa tecnica come acceleratore dell'apprendimento, soprattutto in contesti dove la disponibilità di dati annotati è limitata e la variabilità delle presentazioni cliniche richiede modelli con una forte capacità di generalizzazione. Si evidenzia inoltre, come l'utilizzo di un modello MSA pre allenato su dataset più grandi possa mitigare la mancanza di bias induttivi, permettendo loro di apprendere da dataset più ristretti mantenendo un'alta capacità di generalizzazione e mostrando performance competitive.

In conclusione, l'uso di modelli pre-addestrati si rivela quindi una strategia promettente da integrare nella pipeline di addestramento dei modelli pure attention, specialmente in contesti dove la raccolta di grandi quantità di dati annotati è problematica.

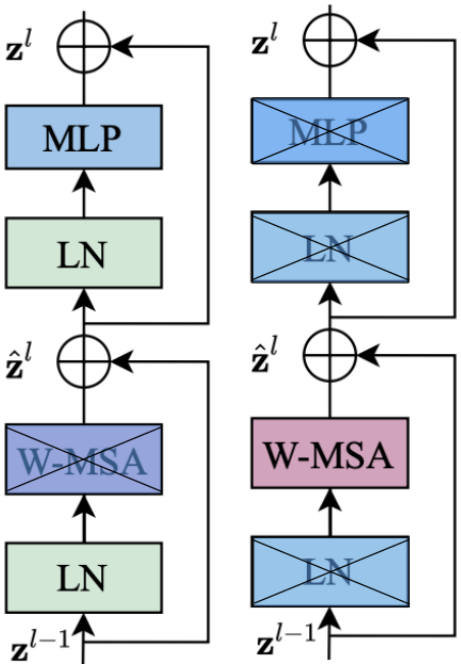
4.3.1 Studio di ablazione sul congelamento dei layer di Swin-T

Il congelamento dei layer è una tecnica comune nel transfer learning, dove gli strati di un modello pre-addestrato sono fissati e solo alcuni strati sono addestrati sul nuovo compito. Per le CNN, questo approccio è motivato dal presupposto che le feature apprese nelle prime fasi di un modello pre-addestrato su grandi dataset (come ImageNet) sono generalmente applicabili a nuovi compiti, mentre l'addestramento degli strati finali permette di adattare il modello alle specificità dei nuovi dati, riducendo anche il requisito computazionale. Nel caso dei modelli transformer, vista la loro natura, ciò potrebbe non essere più valido. Per questo si è deciso di investigare l'impatto del congelamento di alcuni strati del Swin sulle performance. I valori di riferimento per i confronti sono quelli visibili nella tabella 4.3, ottenuti dal Swin senza layer bloccati e con solo la modifica della testa di classificazione.

Prendendo come riferimento [34], in cui è stata svolta un indagine simile sul ViT, applicando sempre la configurazione vincente, sono stati eseguiti dei cicli di addestramento congelando solo i strati "Feed Forward" oppure solo i strati MSA.

Inoltre, è stato eseguito un ulteriore test andando a congelare i primi quattro Swin Block meno profondi in maniera simile a ciò che viene fatto per le ResNet, nelle quali generalmente il congelamento viene applicato sui blocchi iniziali in quanto sono quelli che apprendono le feature più generali.

Come si evince dalla Tabella 4.4, il congelamento di diversi layer influisce sull'F1-Score e ovviamente sul numero di parametri allenabili. Senza congelare nessun layer, il modello allena tutti i 27.5 milioni di parametri, raggiungendo un F1-Score di 0.6513. Questo risultato viene confrontato con lo scenario in cui si congelano gli strati di attenzione o i feed forward, che vedono una riduzione di parametri a 18.9 milioni e 10.2 milioni (meno della metà dei 25 M di ResNet50) rispettivamente.



(a) MSA (b) Feed Forward

Figura 4.5: Congelamento dei moduli MSA e Feed Forward

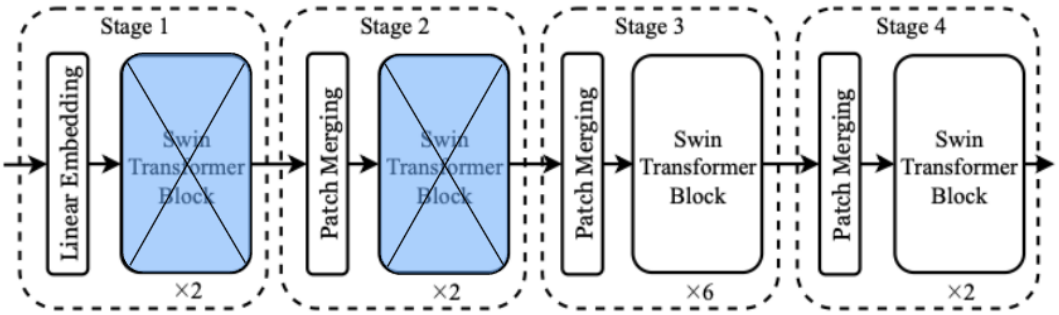


Figura 4.6: Congelamento dei primi 4 Swin Block

Nel caso del blocco dei soli MSA, si nota a grande sorpresa un calo di performance in termini di F1-score del solo 1% con circa 8 M di pesi in meno da aggiornare. Tuttavia è interessante notare come il congelamento dei feed forward, che costituiscono una parte importante della rete in termini di dimensioni, porti delle performance paragonabili a quelle di riferimento con un modello decisamente meno complesso da addestrare in termini computazionali. Questo aspetto potrebbe suggerire che delle rappresentazioni efficaci del contesto e delle relazioni tra le parti dell'immagine generate dai MSA, potrebbero essere più importanti per la performance che non la capacità di elaborazione delle attention map stesse da parte dei feed forward. Un seconda ipotesi potrebbe essere che la riduzione significativa dei parametri abbia avuto un effetto positivo sulla capacità generalizzativa. Le specifiche dinamiche di questo comportamento quindi, non sono ben chiare e possono essere oggetto di studi approfonditi per una migliore comprensione dei meccanismi interni dei Transformer nel trattamento delle attention map. D'altro canto, il congelamento dei blocchi Swin iniziali non offre vantaggi significativi. Le performance raggiunte sono le peggiori in confronto con gli altri due metodi e il risparmio computazionale è minimo.

In conclusione, questo studio di ablazione offre spunti preziosi sul ruolo dei diversi strati nell'adattamento a nuovi domini e sottolinea l'importanza di un'attenta selezione delle parti da congelare per bilanciare la complessità e le prestazioni.

Layer Congelati	**F1-Score**	**Accuracy**	**AUROC**	**Parametri allenabili**
Nessun Layer	0.6513	0.6594	0.8311	27.5 M
Attention	0.6425	0.6454	0.8241	18.9 M
Feed Forward	**0.6503**	**0.6497**	**0.8283**	**10.2 M**
Swin Block 1-4	0.6387	0.6433	0.8268	26.3 M

Tabella 4.4: Performance Swin-T in base ai strati congelati

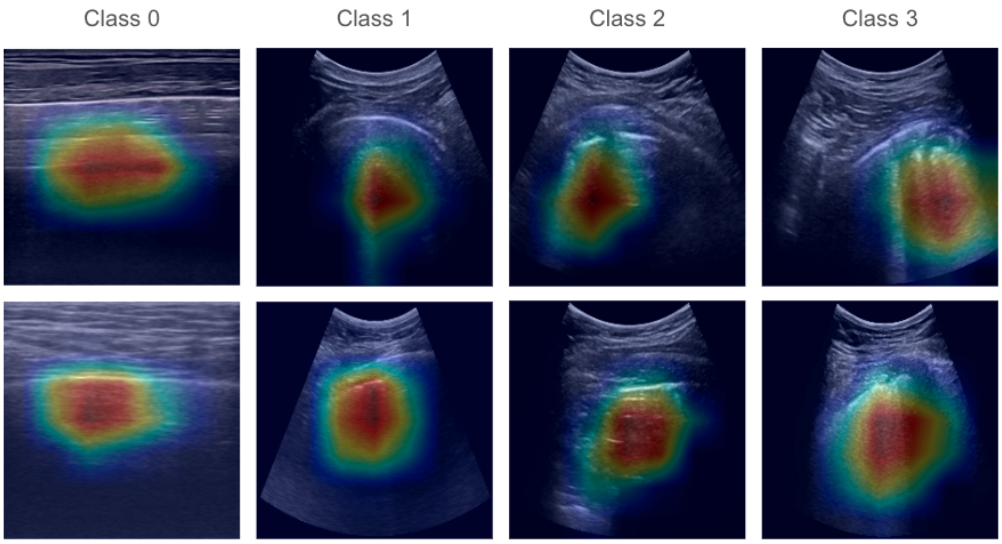
4.3.2 GradCAM

Grad-CAM (Gradient-weighted Class Activation Mapping) è una tecnica di visualizzazione che consente l'interpretazione dei modelli di deep learning attraverso la localizzazione delle regioni d'interesse nelle immagini. Per una data predizione di una classe da parte del modello, tramite Grad-CAM viene generata una mappa di calore che evidenzia le aree hanno influenzato maggiormente la predizione. Nel contesto della diagnostica medica, l'utilizzo di queste mappe di calore aiuta a validare la focalizzazione del modello sulle strutture anatomiche rilevanti, corrispondenti alle caratteristiche patologiche visualizzate nelle immagini ecografiche LUS. La capacità di questi modelli di porre l'attenzione sulle parti corrette delle immagini è fondamentale, in quanto fornisce una conferma che il processo di apprendimento si allinea con la conoscenza medica e le aspettative cliniche. Questa sezione si dedica a un confronto approfondito delle mappe di calore generate dai modelli ResNet50 e Swin Tiny

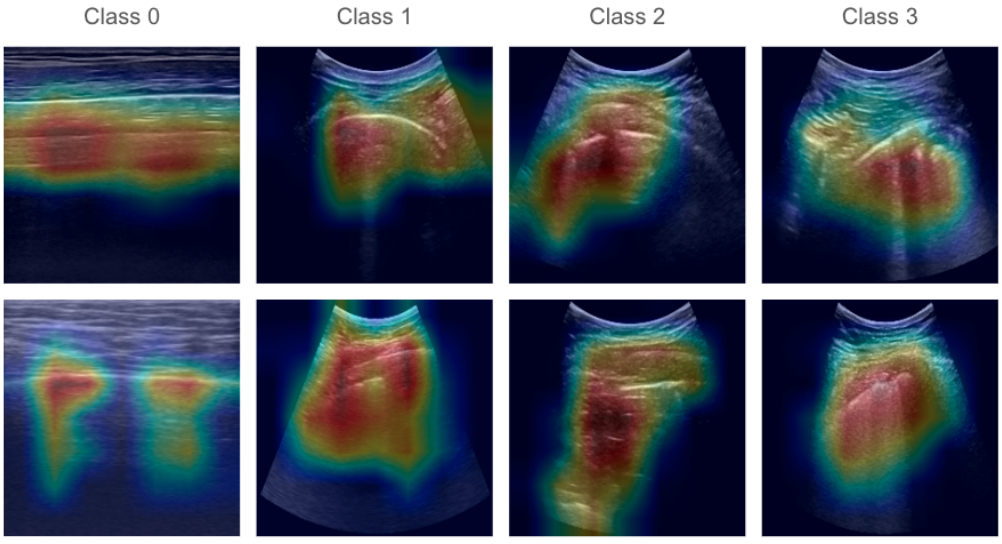
pre-addestrati su ImageNet-1k dopo l'adattamento sul dataset ICLUS (Figura 4.7), focalizzandosi su come identificano e classificano le caratteristiche chiave associate a varie condizioni polmonari. È importante sottolineare che le mappe di calore sono state generate su frame non utilizzati nel processo di addestramento. Vengono presentate due mappe per ciascuna classe del problema (descritte nel dettaglio in 3.1.3)

Osservando la Figura 4.7a ResNet50 mostra una tendenza a concentrarsi sulle texture e sui contorni locali. Le mappe di calore per la classe 0 evidenziano una chiara attenzione alle linee pleuriche, anche se limitate in una area circoscritta. Per le classi patologiche, la focalizzazione si distribuisce su artefatti verticali e polmoni bianchi, ma con un livello di precisione non particolarmente elevato. Per la classe 1 il modello sembra in grado di individuare correttamente le B-line ma le attivazioni presentano una scarsa intensità sulla parte superiore dell'immagine e sulla linea pleurica. Questa tendenza sembra ripresentarsi anche per le classi 2 e 3. Osservando le rispettive mappe si può notare come le attivazioni siano abbastanza intense nelle zone sottostanti della linea pleurica per la presenza di B-line e polmone bianco ma poco intense o addirittura assenti nelle regioni della pleura dove sono presenti le interruzioni o i consolidamenti. Questo suggerisce che la ResNet50 classifica i frame basandosi principalmente sulla presenza di artefatti come B-Line e white lung, nella presenza dei quali ignora in parte o completamente la regione pleurica.

D'altra parte, Swin Tiny, presenta mappe di calore che suggeriscono un'elaborazione dell'immagine più olistica. Come si può osservare nella Figura 4.7b, le aree di attenzione sono più diffuse e si estendono oltre i confini locali degli artefatti sui cui la ResNet50 si focalizza. Ad esempio, nelle immagini corrispondenti alla classe 1, dove devono essere identificate le B-line, Swin Tiny non solo rileva queste formazioni ma sembra anche valutare la continuità della linea pleurica. Analogamente, nelle classi 2 e 3, oltre che a riconoscere la presenza del polmone bianco e artefatti verticali, valuta anche l'estensione di queste anomalie lungo la pleura per identificare interruzioni e consolidamenti. Confrontando direttamente le Grad-CAM delle due architetture, si osserva quindi che Swin Tiny sembra avere una maggiore sensibilità nella rilevazione delle strutture complesse rispetto a ResNet50. La mappe di calore di ResNet50 tendono a concentrarsi su aree più limitate, suggerendo una focalizzazione su artefatti ben specifici. Nel caso di Swin, le mappe di calore evidenziano una maggiore tendenza a includere contesti più ampi nell'analisi, potenzialmente permettendo al modello di catturare le relazioni spaziali tra artefatti polmonari diversi, come consolidamenti e zone di polmone bianco. Questo approccio potrebbe essere particolarmente vantaggioso per identificare le classi con condizioni più severe, dove una comprensione olistica della struttura polmonare è essenziale.



(a) ResNet50



(b) Swin Tiny

Figura 4.7: Mappe di attivazione GradCAM di ResNet50 e Swin Tiny pre-allenati

Capitolo 5

Conclusioni

Questa tesi ha avuto come obiettivo principale l'analisi comparativa dell'applicabilità dei modelli basati su Transformer rispetto alle Reti Neurali Convoluzionali (CNN) nella classificazione delle immagini ecografiche LUS. In particolare, si è cercato di valutare se l'impiego di meccanismi di Multi-Head Self Attention (MSA) potesse offrire un'alternativa valida alle CNN nell'ambito specifico dell'imaging polmonare, focalizzandosi sulla capacità di classificare correttamente casi di particolare criticità e sulla generalizzabilità dei modelli. La ricerca ha esaminato le performance di approcci puramente basati su MSA contro quelli ibridi che combinano MSA e convoluzioni. Si è cercato di stabilire quale tra questi risulta più efficace, soprattutto alla luce delle sfide specifiche del dominio ecografico, tra cui la scarsità e la complessità del dataset ICLUS.

Questa tesi ha esplorato l'applicabilità dei modelli basati su Transformer nella classificazione delle ecografie polmonari LUS in confronto con ResNet50. In particolare sono state esaminate due architetture:

- Swin, un modello di *pure attention* basato esclusivamente su MSA senza convoluzioni, efficace nell'analizzare relazioni a lungo raggio, introducendo un approccio gerarchico efficiente e scalabile basato su MSA calcolata su finestre scorrevoli;
- BoTNet50, un modello ibrido nel quale la MSA viene introdotta semplicemente sostituendo le convoluzioni nei strati più profondi per combinare in efficientemente convoluzioni e self-attention, migliorando le prestazioni nella classificazione grazie alla capacità di catturare caratteristiche globali e locali.

L'obiettivo era determinare quale approccio fosse più efficace nel contesto specifico dell'imaging polmonare, considerando la capacità di classificare correttamente casi critici e la generalizzabilità dei modelli, data la scarsità e la complessità del dataset ICLUS.

Sono state affrontate diverse sfide metodologiche. Una delle più significative è stata la gestione dello sbilanciamento delle classi, affrontata attraverso l'implementazione di una funzione di loss pesata e di metriche più affidabili. L'adozione di scheduler come il Cosine Annealing ha contribuito a migliorare la convergenza dei modelli durante l'addestramento. Inoltre, sono state applicate tecniche di regolarizzazione

e data augmentation per contrastare l'overfitting e incrementare la robustezza dei modelli. Infine, la strategia di Grid Search è stata utilizzata per ottimizzare gli iperparametri, assicurando che i modelli fossero testati con la configurazione più efficace.

Il modello ibrido BoTNet50, che integra il meccanismo di Multi-Head Self Attention (MSA) con una ResNet50 sostituendo gli ultimi strati convoluzionali, ha dimostrato prestazioni superiori e una maggiore solidità rispetto alla tradizionale ResNet50 nella classificazione delle immagini ICLUS, superandola non solo in termini di F1-score ma anche in termini di efficienza grazie ad un numero inferiore di parametri. L'aspetto innovativo del BoTNet50 risiede appunto nella sua capacità di sfruttare in maniera efficiente la MSA globale applicandola sulle feature map convoluzionali caratterizzate da una complessità spaziale inferiore.

Swin Tiny ha invece mostrato delle difficoltà nel raggiungere performance competitive quando addestrato da zero. La natura del modello, che si avvale solo dell'attenzione per catturare le dipendenze globali, si contrappone alla capacità delle CNN di sfruttare i bias induttivi delle convoluzioni. L'addestramento dei Transformer senza un ampio dataset che permetta di apprendere tali dipendenze può quindi risultare problematico, e questo è stato un fattore chiave nei risultati osservati. La limitata quantità di dati si è rivelata un ostacolo significativo per Swin Tiny, il che sottolinea l'importanza di strategie come il transfer learning per compensare la mancanza di bias induttivi che sono naturalmente presenti nelle CNN. I risultati ottenuti applicando questo approccio hanno confermato l'efficacia nel adattare i pesi del modello pre-allenato al dataset specializzato, evidenziando un notevole miglioramento per Swin-T in termini sia di accuratezza che di F1-Score. Il modello è riuscito a ottenere performance competitive superando la ResNet50 in tutte le metriche, colmando il divario prestazionale.

Dallo studio di ablazione sul congelamento di vari strati del Swin è emerso che bloccare l'adattamento dei blocchi iniziali non comporta nessun vantaggio significativo in termini di complessità in rapporto al risultante peggioramento. La situazione cambia quando si blocca l'addestramento degli strati feed forward. Nonostante il congelamento di questi strati che compongono una parte importante della rete in termini di dimensioni, sono state raggiunte prestazioni paragonabili a quelle ottenute con un modello completamente addestrato, però con una complessità computazionale significativamente ridotta. Il fatto che il congelamento di questi strati non abbia diminuito in modo significativo le prestazioni suggerisce che le rappresentazioni del contesto e delle relazioni tra le diverse parti dell'immagine generate dai MSA siano più influenti per la performance del modello rispetto alla capacità di elaborazione dei feed forward. Un'altra ipotesi potrebbe essere la riduzione significativa dei parametri risultante dal congelamento dei feed forward abbia avuto un effetto positivo sulla capacità generalizzativa del modello.

Nell'analisi delle Grad-CAM dei modelli pre-addestrati, Swin Tiny ha dimostrato una sensibilità superiore rispetto a ResNet50, non solo nella rilevazione di specifiche

formazioni, come le B-line, ma anche nel valutare dettagli cruciali nella parte superiore delle immagini. La mappe di calore di ResNet50 tendono a concentrarsi su aree più limitate, suggerendo una focalizzazione su artefatti ben specifici da parte del modello per effettuare la predizione. Nel caso di Swin, le mappe di calore evidenziano la tendenza a includere contesti più ampi nell'analisi, potenzialmente permettendo al modello di catturare le relazioni spaziali tra artefatti polmonari diversi, come consolidamenti e continuità della linea pleurica.

In conclusione, questa tesi ha come contributo l'esplorazione dell'applicabilità dei modelli che sfruttano il meccanismo MSA nel trattamento delle ecografie LUS. BoTNet50 ha mostrato potenziale come backbone con MSA superando in modo consistente ResNet50, mentre il modello pure attention Swin Tiny non è risultato all'altezza del task quando allenato da zero. Un ulteriore contributo è dato dall'applicazione di metodi del transfer learning per compensare alla mancanza di bias induttivi nel Swin, ottenendo risultati competitivi con lo stato dell'arte. Inoltre, grazie allo studio di ablazione si è compreso che è possibile raggiungere performance simili riducendo notevolmente la complessità del modello durante l'addestramento bloccando particolari strati. Infine, grazie alle Grad-CAM si è potuto osservare la capacita del Swin di considerare attivazioni più diversificate per la predizione delle classi.

5.1 Limitazioni dello studio

Come già sottolineato più volte in precedenza, questo studio non è esente da limitazioni significative. In primo luogo, non è stato possibile svolgere un confronto approfondito e accurato con lo stato dell'arte a causa della indisponibilità dei hold-out utilizzati nei lavori di riferimento. Inoltre, la mancanza di modelli transformer per il trattamento delle ecografie LUS rende questa ricerca la prima ad applicare tali metodologie al dataset ICLUS. A questo bisogna considerare anche le difficoltà intrinseche nel trattare i dataset clinici. Quindi, seppur i risultati ottenuti indichino le potenzialità della MSA per questo problema, soprattutto nel caso del Swin pre-allenato, sono necessarie ulteriori convalide e confronti con altre metodologie presenti in letteratura, anche adottando altre backbone convoluzionali di riferimento oltre a ResNet50. Altri limiti derivano dalla variabilità dei risultati in base alle diverse possibilità di splitting dei pazienti nei vari set portando alla necessita di valutare l'impatto di questo fenomeno e di applicare approcci per limitarlo.

5.2 Sviluppi futuri

In futuro, ulteriori ricerche potrebbero concentrarsi sullo sviluppo di metodi più efficaci per la gestione delle limitazioni dei dati e sull'esplorazione di nuovi modelli di Vision Transformer, potenzialmente ampliando il campo di applicazione a diverse

modalità di imaging medico. Un altro possibile sviluppo è considerare l'integrazione dei bias induttivi specifici al dominio medico nei modelli transformer pure attention.

Di possibile interesse futuro riguarda l'utilizzo delle metodologie di splitting dei lavori di riferimento, che potrebbe consentire un confronto più approfondito con i risultati attuali. Inoltre, l'adozione di tecniche come la cross-validation potrebbe fornire un approccio più robusto e affidabile nell'analisi delle prestazioni.

Ulteriori percorsi di sviluppo includono l'integrazione di informazioni di dominio nel processo di classificazione, seguendo l'esempio della ricerca di Frank et al.[9], l'integrazione di approcci multi-task e l'adozione di metodi ordinali. Questi sviluppi potrebbero fornire un framework più robusto e versatile per la classificazione e l'analisi delle immagini LUS e un ulteriore passo avanti nello stato dell'arte.

Bibliografia

[1] Aravind Srinivas, Tsung Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2021.

[2] Fabio Fichera, Manuela Nicotra, and Italo Paolini. Utilità della pocus toracica in corso di infezione da COVID-19. *Rivista Società Italiana di Medicina Generale n. 2*, 27, 2020.

[3] Antonello D'Andrea, Giovanna Di Giannuario, Gemma Marrazzo, Lucia Riegler, Donato Mele, Massimiliano Rizzo, Marco Campana, Alessia Gimelli, Georgette Khoury, and Antonella Moreo. L'imaging integrato nel percorso del paziente con COVID-19: dalla diagnosi, al monitoraggio clinico, alla prognosi. *Giornale Italiano di Cardiologia n. 5*, 21, 2020.

[4] Andrea Genovese. Ecografia e COVID-19: quale ruolo? `https://www.h3-surgical-team.it/blog/ecografia-e-covid19/`, 2020.

[5] Lumin Xing, Wenjian Liu, Xiaoliang Liu, and Xin Li. An enhanced vision transformer model in digital twins powered internet of medical things for pneumonia diagnosis, 2023.

[6] Oğuzhan Katar, Ozal Yildirim, and Yeşim Eroğlu. Vision transformer model for efficient stroke detection in neuroimaging. *2023 4th International Informatics and Software Engineering Conference (IISEC)*, pages 1–6, 2023.

[7] S. Tummala, Seifedine Kadry, Syed Ahmad Chan Bukhari, and Hafiz Tayyab Rauf. Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. *Current Oncology*, 29:7498 – 7511, 2022.

[8] Subhankar Roy, Willi Menapace, Sebastiaan Oei, Ben Luijten, Enrico Fini, Cristiano Saltori, Iris Huijben, Nishith Chennakeshava, Federico Mento, Alessandro Sentelli, Emanuele Peschiera, Riccardo Trevisan, Giovanni Maschietto, Elena Torri, Riccardo Inchingolo, Andrea Smargiassi, Gino Soldati, Paolo Rota, Andrea Passerini, Ruud J.G. Van Sloun, Elisa Ricci, and Libertario Demi. Deep learning for classification and localization of covid-19 markers in point-of-care lung ultrasound. *IEEE Transactions on Medical Imaging*, 39, 2020.

[9] Oz Frank, Nir Schipper, Mordehay Vaturi, Gino Soldati, Andrea Smargiassi, Riccardo Inchingolo, Elena Torri, Tiziano Perrone, Federico Mento, Libertario Demi, Meirav Galun, and Yonina C. Eldar. Integrating domain knowledge into deep networks for lung ultrasound with applications to covid-19. *IEEE Transactions on Medical Imaging*, 41, 2022.

[10] Umair Khan, Sajjad Afrakhteh, Federico Mento, Noreen Fatima, Laura De Rosa, Leonardo Lucio Custode, Zihadul Azam, Elena Torri, Gino Soldati, Francesco Tursi, Veronica Narvena Macioce, Andrea Smargiassi, Riccardo Inchingolo, Tiziano Perrone, Giovanni Iacca, and Libertario Demi. Benchmark methodological approach for the application of artificial intelligence to lung ultrasound data from covid-19 patients: From frame to prognostic-level. *Ultrasonics*, 132, 2023.

[11] Szymon Płotka, Michal K. Grzeszczyk, Robert Brawura-Biskupski-Samaha, Paweł Gutaj, Michał Lipa, Tomasz Trzciński, and Arkadiusz Sitek. Babynet: Residual transformer module for birth weight prediction on fetal ultrasound video. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13434 LNCS, 2022.

[12] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann Lecun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018.

[13] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2, 2018.

[14] Behnaz Gheflati and Hassan Rivaz. Vision transformers for classification of breast ultrasound images. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2022-July, 2022.

[15] Irwan Bello, Barret Zoph, Quoc Le, Ashish Vaswani, and Jonathon Shlens. Attention augmented convolutional networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-October, 2019.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR 2021 - 9th International Conference on Learning Representations*, 2021.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-December, 2017.

[18] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. (ln) layer norm. *arXiv:1607.06450v1*, 2015.

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[20] Gino Soldati, Andrea Smargiassi, Riccardo Inchingolo, Danilo Buonsenso, Tiziano Perrone, Domenica Federica Briganti, Stefano Perlini, Elena Torri, Alberto Mariani, Elisa Eleonora Mossolani, Francesco Tursi, Federico Mento, and Libertario Demi. Proposal for international standardization of the use of lung ultrasound for patients with covid-19. *Journal of Ultrasound in Medicine*, 39, 2020.

[21] Gino Soldati, Andrea Smargiassi, Riccardo Inchingolo, Danilo Buonsenso, Tiziano Perrone, Domenica Federica Briganti, Stefano Perlini, Elena Torri, Alberto Mariani, Elisa Eleonora Mossolani, Francesco Tursi, Federico Mento, and Libertario Demi. Is there a role for lung ultrasound during the covid-19 pandemic?, 2020.

[22] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. pages 464–468, 2018.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December, 2016.

[24] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. Attention augmented convolutional networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3285–3294, 2019.

[25] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models, 2019.

[26] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks, 2017.

[27] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning, 17–19 Jun 2013.

[28] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks, 2017.

[29] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017.

[30] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get m for free, 2017.

[31] Yunsung Lee, Gyuseong Lee, Kwang seok Ryoo, Hyojun Go, Jihye Park, and Seung Wook Kim. Towards flexible inductive bias via progressive reparameterization scheduling. pages 706–720, 2022.

[32] Yuxuan Zhou, Wangmeng Xiang, C. Li, Biao Wang, Xihan Wei, Lei Zhang, M. Keuper, and Xia Hua. Sp-vit: Learning 2d spatial priors for vision transformers. 2022.

[33] Chongzhi Zhang, Mingyuan Zhang, Shanghang Zhang, Daisheng Jin, Qiang-feng Zhou, Zhongang Cai, Haiyu Zhao, Shuai Yi, Xianglong Liu, and Ziwei Liu. Delving deep into the generalization of vision transformers under distribution shifts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[34] Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Jakob Verbeek, and Hervé Jégou. Three things everyone should know about vision transformers, 2022.