Semi-Supervised Lung Nodule Malignancy Classification Using Visual Attribute Pseudo-Labels

José Pedro Figueiredo Machado
Master in Data Science
Department of Computer Science
Faculty of Sciences of the University of Porto
2025-2026

Semi-Supervised Lung Nodule Malignancy Classification Using Visual Attribute Pseudo-Labels

José Pedro Figueiredo Machado

Dissertation carried out as part of
the Master in Data Science
Department of Computer Science
2025-2026

Supervisor
Hélder Oliveira, Associate Professor, FCUP, INESC-TEC

Co-supervisor
Tânia Pereira, Assistant Professor, FEUP, INESC-TEC

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

INESCTEC

Acknowledgements

First, I would like to thank my advisor, Hélder Oliveira, for the opportunity to be part of this project and believing in me.

To Eduardo Rodrigues and Tânia Pereira, for their support and willingness to help, for putting themselves in my shoes, and for advising me with genuine concern and care.

To my group of friends, for getting me out of the house and helping me break out of my routine. For making my journey lighter.

To Ana, who believed in me when I couldn't and taught me to be kinder to myself. Your warm personality, constant concern, and care made me realize that what I do is never just for myself.

To my family, for being the foundation of everything, for your unconditional support and constant presence.

To Mia, whose snoring has become the sound I work best to.

This work was funded by the European Union under the Horizon Europe Programme (Grant Agreement No. 101080756). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

Agradecimentos

Primeiramente, queria agradecer ao meu orientador, Hélder Oliveira, pela oportunidade de fazer parte deste projecto e por me ter acreditado capaz.

Ao Eduardo Rodrigues e Tânia Pereira, pelo apoio e disponibilidade, por se saberem por no meu lugar e me aconselharem com preocupação e cuidado genuíno.

Ao meu grupo de amigos, por me tirarem de casa e me ajudarem a sair da rotina. Por tornarem mais leve o meu caminho.

À Ana, que acreditou em mim quando eu não conseguia e me ensinou a ser mais gentil comigo mesmo. A tua personalidade calorosa, preocupação e cuidado constantes fizeram-me perceber que aquilo que faço nunca é apenas por mim.

À minha família, por serem a base de tudo, pelo apoio incondicional e presença constante.

À Mia, cujo ressonar se tornou o som ao qual melhor trabalho.

Este trabalho foi financiado pela União Europeia no âmbito do Programa «Horizon Europe» (Acordo de Subvenção n.º 101080756). As opiniões e pontos de vista aqui expressos são, no entanto, da exclusiva responsabilidade do(s) autor(es) e não refletem necessariamente as da União Europeia ou da Agência Executiva para a Saúde e o Digital (HaDEA). Nem a União Europeia nem a entidade financiadora podem ser responsabilizadas por esses pontos de vista.

Abstract

Faculdade de Ciências da Universidade do Porto
Departamento de Ciência de Computadores

MSc. Data Science

Semi-Supervised Lung Nodule Malignancy Classification Using Visual Attribute Pseudo-Labels

by José Pedro Figueiredo Machado

Lung cancer remains one of the leading causes of cancer-related mortality worldwide, largely due to late-stage diagnosis and the difficulty of characterizing pulmonary nodules from computed tomography imaging. While deep learning has shown strong potential for automated malignancy classification, most existing approaches disregard the clinically meaningful visual attributes routinely used by radiologists in diagnostic decision-making. This dissertation proposes a unified framework that jointly optimizes malignancy classification and lung nodule visual attributes prediction through multi-task learning, further extended with semi-supervised consistency regularization via FixMatch and UniMatch, combining the LIDC-IDRI and LUNA25 datasets. Multi-task learning with LNVA supervision matches or slightly exceeds the supervised baseline on malignancy classification, confirming that visual attribute annotations provide a compatible inductive bias for the primary task. The semi-supervised extension produced limited gains, with pseudo-label utilization below 15% across all configurations, suggesting the bottleneck lies in generating reliable pseudo-labels for multi-class ordinal targets under domain shift rather than in the absence of useful unlabeled data. Explainability analysis via Grad-CAM further reveals that attribute-guided supervision shifts spatial attention toward boundary and peri-nodular regions in a manner anatomically consistent with the attributes, indicating that multi-task learning shapes a more clinically grounded internal representation beyond what malignancy supervision alone achieves.

Resumo

Faculdade de Ciências da Universidade do Porto
Departamento de Ciência de Computadores

Mestrado em Ciências de Dados

Classificação semi-supervisionada de malignidade em nódulos pulmonares através atributos visuais

por José Pedro Figueiredo Machado

Cancro no pulmão continua a ser uma das principais causas de mortalidade oncológica a nível mundial, em grande parte devido ao diagnóstico tardio e à dificuldade em caracterizar nódulos a partir de imagens de tomografia. Embora *deep learning* tenha demonstrado potencial para a classificação de malignidade, a maioria das abordagens existentes ignora os atributos visuais clinicamente relevantes que os radiologistas utilizam rotineiramente na tomada de decisão. Esta tese propõe uma estrutura que classifica malignidade e prevê atributos visuais simultaneamente através de *multi-task learning*. Estendido depois a *semi-supervised learning* via FixMatch e UniMatch, combinando os *datasets* LIDC-IDRI e LUNA25. Neste trabalho, *multi-task learning* com supervisão de atributos visuais iguala ou supera a classificação malignidade base, confirmando que as anotações de atributos visuais constituem, no minimo, um bias indutivo compatível com a tarefa principal. A extensão a *semi-supervised* produziu ganhos limitados, com uma taxa de utilização de pseudo-rótulos inferior a 15% em todas as configurações, sugerindo que o principal obstáculo reside na geração de pseudo-rótulos fiáveis, e não aa ausência de dados não rotulados úteis. A análise de interpretabilidade via Grad-CAM revela ainda que a supervisão guiada por atributos desloca a atenção espacial do modelo para regiões de fronteira e peri-nodulares de forma consistente com os atributos, indicando que *multi-task learning* constrói uma representação interna mais clinicamente fundamentada do que a supervisão de malignidade isolada consegue alcançar.

Contents

List of Figures

List of Tables

Symbols

λ	loss weighting factor (malignancy, LNVA or SSL term)	–
τ	confidence threshold (FixMatch/UniMatch)	–
α	semi-supervised consistency loss weight	–
γ	focusing parameter, Focal Loss	–
w_a	disagreement-aware reliability weight for attribute a	–
H_a	normalized entropy of attribute a	–
$\hat{y}$	pseudo-label	–
x^w	weakly augmented view of an input sample	–
x^s	strongly augmented view of an input sample	–
z	latent representation, $z = g(.)$	–
m	confidence mask	–
S_a	attribute importance score (logistic regression coefficient norm)	–

Glossary

ACRIN	American College of Radiology Imaging Network
AUROC	Area Under the Receiver Operating Characteristic Curve
BB	Bounding Box
BCE	Binary Cross-Entropy
BN	Batch Normalization
CAD	Computer-Aided Diagnosis
CE	Cross-Entropy
CNN	Convolutional Neural Network
CT	Computed Tomography
DCE	Disagreement-aware Cross-Entropy
DWCE	Disagreement Weighted Cross-Entropy
FDA	Food and Drug Administration
FL	Focal Loss
FN	False Negative
FNIH	Foundation for the National Institutes of Health
FP	False Positive
GAN	Generative Adversarial Network
GGO	Ground-glass Opacity

Grad-CAM	Gradient-weighted Class Activation Mapping
HSCNN	Hierarchical Semantic CNN
LIDC-IDRI	Lung Image Database Consortium and Image Database Resource Initiative
LNDb	Lung Nodule Database
LNVA	Lung Nodule Visual Attribute
LR	Learning Rate
LSS	Lung Screening Study group
NCI	National Cancer Institute
NLST	National Lung Screening Trial
SLURM	Simple Linux Utility for Resource Management
SSL	Semi-Supervised Learning
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
WCE	Weighted Cross-Entropy

Chapter 1

Introduction

1.1 Context

Lung cancer remains one of the leading causes of cancer-related mortality worldwide [1], largely due to its late-stage diagnosis and the difficulty of detecting malignancy in early pulmonary stages. Although computed tomography (CT) imaging is a standard tool for lung cancer screening, the interpretation of scans is highly dependent on expert radiologists and is subject to inter-observer variability, particularly in subtle or ambiguous cases [2]. When CT interpretation of professionals is not enough, biopsy and molecular testing remains the best clinical practice for definitive diagnosis [3]. Because these procedures are invasive and associated with potential complications, including pneumothorax and infections [4], Computer-Aided Diagnosis (CAD) systems have been developed to provide non-invasive, data-driven support for lung nodule assessment. This diagnosis often rely on deep learning methods, that have demonstrated strong potential in learning discriminative representations directly from CT imaging data [5]. However, their effectiveness is often limited by the scarcity of high-quality annotated datasets and the lack of structured supervision that reflects clinically meaningful nodule characteristics beyond binary malignancy labels [6].

1.2 Motivation

Most existing deep learning approaches for lung nodule analysis focus exclusively on malignancy classification, disregarding intermediate radiological descriptors that are routinely used in clinical decision-making [7], limiting both the interpretability and generalization of the learned representations. At the same time, available datasets present heterogeneous characteristics: LIDC-IDRI provides detailed malignancy and visual attributes annotations, while LUNA25 offers a larger set of nodules, with limited but precise labeling. This imbalance motivates the development of learning strategies capable of jointly exploiting multiple sources of supervision.

1.3 Objectives

The primary objective of this thesis is to design and evaluate a unified deep learning framework for pulmonary nodule analysis that integrates malignancy and lung nodule visual attributes prediction thought semi-supervised learning. More specifically, this work aims to evaluate the impact of multi-task learning on malignancy classification through auxiliary lung nodule visual attributes (LNVA) supervision and investigate the contribution of semi-supervised consistency regularization to that purpose. All of this while providing a controlled experimental analysis that isolates the effect of each component of the proposed framework, enabling a systematic comparison of multi-task learning strategies, semi-supervised learning formulations, and convolutional backbone architectures.

1.4 Thesis Structure

This thesis is organized into six chapters, each addressing a distinct stage of the proposed framework.

Chapter 2 reviews the relevant literature across four threads: semi-supervised learning in medical imaging, deep learning approaches to lung nodule malignancy classification, semi-supervised methods applied specifically to lung nodules and attribute-guided learning strategies.

Chapter 3 formalizes the problem and describes the proposed framework. It introduces

the two datasets used throughout this work, LIDC-IDRI and LUNA25, detailing the multi-task learning setup, proposed architecture and defines the training objective. Finally the staged experimental design strategy that structures the remaining chapters is presented.

Chapter 4 describes the practical conditions under which all experiments were conducted, covering the computational environment, software stack, data preprocessing pipeline, loss functions, semi-supervised augmentation strategies, experimental configuration and evaluation protocol.

Chapter 5 reports and discusses the results across the four experimental phases: the supervised baseline, multi-task learning, semi-supervised learning and interpretability analysis.

Chapter 6 summarizes the main findings with respect to the three research questions, discusses the limitations of the proposed framework and outlines directions for future work.

Chapter 2

State of the Art

Lung nodule analysis has been extensively studied using classical supervised classification approaches, with increasing attention given to semi-supervised frameworks that leverage unlabeled data. This chapter reviews the most relevant developments in each of these threads: section 2.1 introduces semi-supervised learning in the broader context of medical imaging and section 2.2 surveys deep learning approaches to lung nodule malignancy classification, with particular attention to work on the LIDC-IDRI benchmark. Section 2.3 narrows the focus to semi-supervised methods applied specifically to lung nodule analysis, and includes a comparative overview of representative works. Finally, Section 2.4 reviews attribute-guided learning approaches, tracing the line from early multi-task CNNs to the most recent prototype-based interpretable architectures.

2.1 Semi-supervised learning in medical imaging

Most of the success in the application of deep learning in medical imaging has a strong dependence on the availability of large-scale annotated datasets. These are expensive and difficult of obtain due to the need for the field knowledge from experts and the time-consuming nature of the labeling process. In contrast, there are large amounts of unlabeled medical data available, and that is where the semi-supervised learning approach finds its place, leveraging both labeled and unlabeled data while improving the model performance. Among the numerous semi-supervised learning (SSL) strategies, consistency regularization*, often combined with pseudo-labeling mechanisms†, emerged as the dominant approach. In recent works, specially in medical image segmentation, it shows

*Learning strategy that enforces consistent predictions under input perturbations.
†Approach where high-confidence predictions on unlabeled data are used as targets for further training.

that enforcing prediction consistency under input perturbations can significantly improve performance when labeled data is limited [5].

2.2 Lung nodule malignancy classification

Lung nodule malignancy classification is a critical step in computer-aided diagnosis systems, but also inherently challenging due to the high variability in nodule appearance and the subjective nature of radiologist annotations. A comprehensive systematic review by Lea Marie Pehrson *et al.* [7] analyzed 41 studies that apply machine learning and deep learning methods to the LIDC-IDRI dataset. Feature-based methods that typically rely on a combination of handcrafted descriptors and classifiers like Support Vector Machines (SVM) can achieve the 99% accuracy mark in some cases. On the other side, Deep learning approaches, particularly convolutional neural networks (CNNs), demonstrated strong performance with accuracies reported ranging from 82.2% to 97.6%, while benefiting from automatic feature extraction and the reduced reliance on manual preprocessing. Despite these promising results, the review highlights a lack of standardization in evaluation protocols, making direct comparison between methods difficult.

Several architectural innovations have been proposed to address these challenges. Shiwen Shen *et al.* [8] introduced a hierarchical semantic CNN that jointly models low-level features and semantic nodule characteristics in a two-stage process, demonstrating that auxiliary attribute supervision can guide the learning of more discriminative representations. More recently, Xiaohang Fu *et al.* [9] proposed a cross-attention multi-task network that scores all nine LIDC-IDRI attributes simultaneously alongside malignancy, using slice-attention and attribute specialization modules to exploit inter-attribute dependencies. Complementarily, Margarida Gouveia *et al.* [10] addressed a practical limitation common to most benchmark methods: the lack of generalization to out-of-domain data. Using a ResNet backbone with 2.5D inputs and domain-specific data augmentation, they showed consistent performance across both in-domain (LIDC-IDRI, LNDb [11]) and out-of-domain (LUNGx [12]) datasets, underscoring the importance of evaluating generalization as a first-class criterion.

Despite these advances, the limited availability of consistent labeled data remains a major

bottleneck. These limitations motivate the exploration of approaches like semi-supervised learning, that can better leverage available data.

2.3 Semi-supervised learning applied to lung nodules

Early approaches to SSL, in this context, are primarily based on pseudo-labeling mechanisms. For example, Ioannis D. Apostolopoulos *et al.* [13] combined generative adversarial networks (GANs) and a self-training procedure to expand the training dataset by generating synthetic lung nodules. This incorporation of synthetic samples in pseudo-labeling mechanisms showed an improvement in classification performance, but more recent advances aim to add consistency regularization to the mix, like FixMatch [14], that combines consistency constraints with confidence-based pseudo-labeling, achieving strong performance while reducing the reliance on large labeled datasets. This was explored in a recent dissertation by Afonso Esteves Abreu *et al.* [15] where FixMatch was used to classify lung nodule malignancy in a semi-supervised setting. The work combines labeled data from the LIDC-IDRI dataset with unlabeled CT scans from LUNA25, demonstrating that the inclusion of unlabeled data can improve AUROC by approximately 2% compared to a fully supervised baseline. In addition, the study incorporates explainability analysis using Gradient-weighted Class Activation Mapping (Grad-CAM), reporting more localized activation patterns in semi-supervised models, although without definitive clinical interpretability improvements. Despite these gains, the approach remains focused on image-level malignancy classification and does not explicitly incorporate lung nodule visual attributes supervision or a structured multi-task learning framework, which are central to the present work.

Khosravan *et al.* [16] tackled the problem from a multi-task angle, proposing a semi-supervised framework that jointly addresses nodule segmentation and false-positive reduction, demonstrating that combining task diversity with unlabeled data can substantially improve performance beyond either approach in isolation. Wentao Zhu *et al.* [17] extended this direction to a consistency-based framework for nodule classification, enforcing agreement between a student and an exponential-moving-average teacher network under strong perturbations and showing gains over fully supervised baselines with as little as 20% labeled data. More recently, Tushar *et al.* [18] provided a reproducible benchmarking study covering multiple low-dose CT datasets, including the LUNA25 dataset

used in this work, and highlighted the difficulty of cross-dataset generalization, further motivating the use of complementary unlabeled data from heterogeneous sources.

Despite those advances, challenges such as high variability in nodule appearance and annotation uncertainty continue to hinder performance, encouraging the exploitation of consistency-based semi-supervised learning approaches tailored to lung nodule classification. Table 2.1 summarizes the representative SSL works reviewed in this section, highlighting their key design choices.

TABLE 2.1: Comparison of representative semi-supervised learning paradigms for lung nodule analysis.

Work	SSL paradigm	Primary task	Unlabeled signal
Apostolopoulos et al. [13]	GAN + self-training	Malignancy	Synthetic samples
Khosravan & Bagci [16]	Multi-task SSL	Segmentation + FP reduction	Unlabelled CT patches
Zhu et al. [17]	Mean Teacher	Malignancy	Consistency under perturbation
Sohn et al. (FixMatch) [14]	Consistency + pseudo-labeling	General classification	High-confidence predictions
Abreu et al. [15]	FixMatch	Malignancy	LUNA25 unlabeled CT scans
This work	FixMatch + LNVA multi-task	Malignancy + attribute prediction	LUNA25 + LNVA pseudo-labels

2.4 Attribute-guided learning

The explicit modeling of clinically meaningful visual attributes for lung nodule characterization has a history predating the modern deep learning era. Early work by Hancock and Magnan [19] demonstrated that combining radiologist-quantified image features, corresponding closely to the LIDC-IDRI LNVAs, with classical statistical classifiers could yield competitive malignancy prediction, establishing the clinical relevance of these attributes as predictive signals.

The first deep learning work to both model attributes and malignancy was proposed by Sarfaraz Hussein *et al.* [20], who introduced a 3D CNN multi-task learning framework that simultaneously predicts six LIDC-IDRI attributes alongside malignancy by incorporating graph-regularized multi-task loss and transfer learning, demonstrating that attribute supervision acts as a beneficial inductive bias and improving malignancy performance over a single-task baseline. As referenced before, Shiwen Shen *et al.* [8] subsequently introduced the Hierarchical Semantic CNN, in which low-level convolutional

features and higher-level semantic attribute predictions are organized in a two-stage hierarchy, with the final malignancy classification conditioned on both. This showed that structured attribute supervision not only improves accuracy but also provides an interpretable basis for model decisions. A further step towards explainability was taken by Rodney LaLonde *et al.* [21], who proposed X-Caps, the first study to use capsule networks for high-level visual attribute prediction in medical image analysis. By encoding six LIDC-IDRI visual attributes within the capsule vectors of a 2D network, X-Caps demonstrated that interpretable, concept-based representations could approach the malignancy classification performance of non-explainable 3D CNNs.

The most recent work in this line by Eduardo M. Rodrigues *et al.* "Efficient-ProtoCaps" [22], which combines capsule networks with prototype learning for joint malignancy and LNVA prediction on LIDC-IDRI. By introducing a Davies-Bouldin Index with multiple centroids as a clustering loss, the model promotes structured attribute representations while achieving 89.7% accuracy with only approximately 2% of the parameters of the baseline model, providing an inherently interpretable and parameter-efficient architecture. However, the framework is fully supervised and therefore remains dependent on the availability of annotations, which are expensive and difficult to obtain. This direction highlights the importance of attribute-level supervision in lung nodule analysis, beyond purely image-level prediction.

Chapter 3

Methodology

This chapter describes the proposed framework for lung nodule malignancy classification. It formalizes the problem and research questions that guide the experimental design, describes the datasets, the multi-task learning strategies and proposed architecture along side the training objective. A detailed account of the experimental design strategy can be found at the end of the chapter, which structures the progression from supervised learning, through multi-task and ending on semi-supervised extensions to interpretability analysis.

3.1 Problem Formulation

The objective of this work extends beyond conventional malignancy classification. In addition to predicting whether a pulmonary nodule is benign or malignant, the proposed framework aims to learn a set of clinically meaningful lung nodule visual attributes (LNVAs) and investigate their role in both representation learning and model interpretability. Formally, given an input lung nodule patch x, the model jointly predicts malignancy and a set of visual attributes:

$$f(x) \rightarrow \{y_m, y_a\} \tag{1}$$

- y_m represents binary malignancy label
- $y_a = \{a_1, a_2, ..., a_K\}$ represents the set of lung nodule visual attributes

This formulation is motivated by the hypothesis that attribute-guided supervision can

encourage the learning of more structured and clinically meaningful representations than malignancy supervision alone. Consequently, this work investigates three complementary research questions:

1. Can nodule visual attribute supervision improve malignancy classification performance?

2. Can semi-supervised learning improve LNVA prediction by leveraging samples without attribute annotations?

3. Does attribute-guided learning influence the spatial and semantic behavior of the model in an interpretable manner?

3.2 Datasets

LIDC-IDRI and LUNA25 are the two datasets used throughout this work. Together, they support the dual nature of the proposed framework, where the former provides rich radiologist-defined visual attribute annotations alongside malignancy labels, while latter contributes with a substantially larger set of nodules with malignancy labels derived from patient-level outcomes, but without attribute annotations.

3.2.1 LIDC-IDRI

The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) is the dominant benchmark for lung cancer prediction [23]. It was initiated by the National Cancer Institute (NCI), advanced by the Foundation for the National Institutes of Health (FNIH) and accompanied by the Food and Drug Administration (FDA) [6]. Containing 1018 thoracic computed tomography (CT) scans, it went throught a annotation process, performed by four experienced thoracic radiologists, that consisted of an initial blinded-read phase, where each radiologist reviewed the CT scan and classified the lesions as nodules ($\geq 3mm$ and $< 3mm$) and non-nodules ($\geq 3mm$), and then, in a unblinded-read phase each radiologist reviewed their analysis as well as the other professionals in an anonymized manner. This was done with the objective to gather a consensual opinion while maximizing the identification of lung nodules in different CT scans without clashes in reviews. The result was a Database that contains 7371 lesions marked as ”nodule” by at least one radiologist, where 2669 of these were marked ”nodule $\geq$ 3mm”

by at least one radiologist and only 928 (34.7%) of those received such marks from all four radiologists. Because the LIDC-IDRI datasets is a set CT scans collected from multiple institutions, it is important to go through a transformation process to mitigate the variability from the intrinsic differences in CT machines and acquisition protocols. The dataset was already preprocessed with the same algorithm as in "Efficient-ProtoCaps" [22], where the `pylidc` library* was used to read $3D$ array volumes, the annotations and the lung nodule bounding box (BB) coordinates. The preprocessing pipeline includes Hounsfield unit clipping to $[-1000, 400]$, followed by min–max normalization to $[0, 1]$, spatial resampling to an isotropic resolution of $1mm$, and extraction of $2D$ lung nodule-centered patches.

Each nodule is annotated with a set of 9 semantic features, the lung nodule visual attributes (LNVAs), including malignancy, that are scored on a scale detailed in Appendix .1 (Table A.1) and exemplified in the figure bellow Fig 3.1:

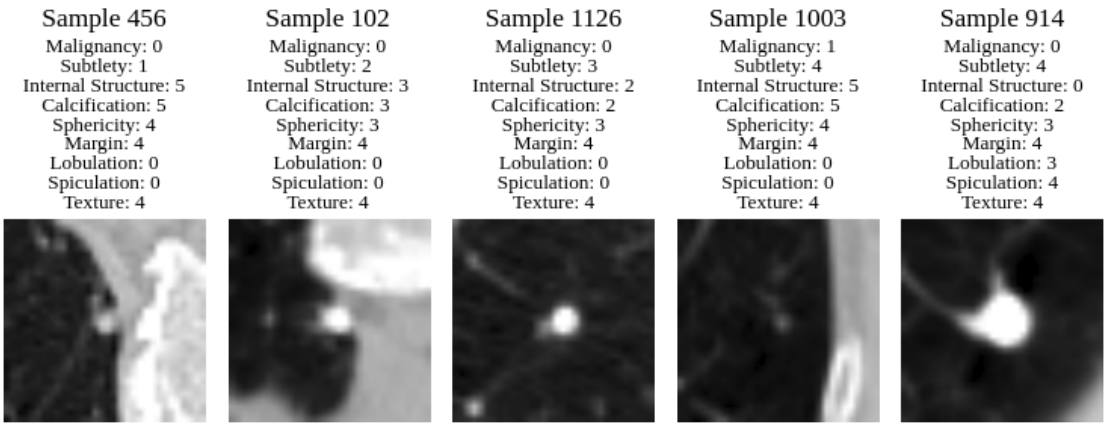


FIGURE 3.1: LIDC-IDRI samples.

3.2.2 LUNA25

LUNA25 is a dataset constructed from the National Lung Screening Trial (NLST), a randomized controlled trial designed to assess differences in lung cancer mortality between low-dose helical computed tomography (CT) [24] screening and chest radiography. Conducted by the Lung Screening Study group (LSS) and the American College of Radiology Imaging Network (ACRIN), the dataset contains approximately $4,000$ CT scans and over $6,000$ annotated nodules [18]. Additionally, it complements LIDC-IDRI in two important ways: first, it contains a substantially larger number of nodules, increasing the diversity

*Library website: https://pylidc.github.io (Accessed on September 15, 2025).

of training samples available; second, its malignancy labels are derived from patient-level outcomes in the NLST rather than subjective radiologist assessments, providing an alternative source of supervision. Its important to note that although this dataset provides malignancy labels, it does not contain radiologist-defined lung nodule visual attributes (LNVA) and consequently, in the semi-supervised stage of this work, it is treated as labeled for malignancy prediction and unlabeled for attribute learning. Example of this images in Fig 3.2.

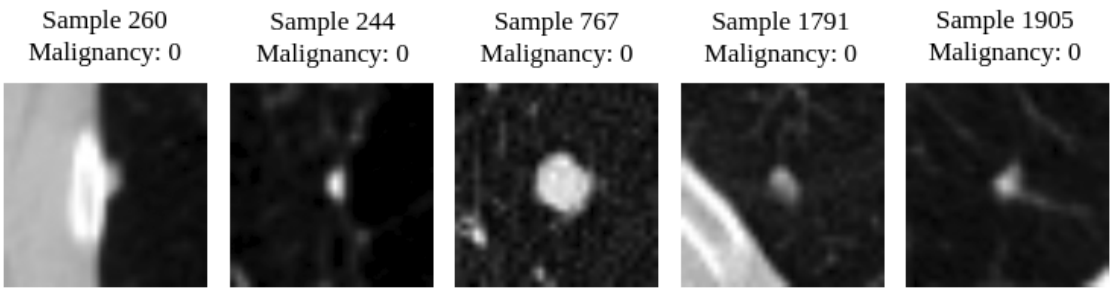


FIGURE 3.2: LUNA25 samples.

3.2.3 Dataset Characteristics

The LIDC-IDRI dataset exhibits substantial inter-observer variability, as each nodule is annotated by multiple radiologists using an ordinal scale. This variability introduces inherent label uncertainty, which is reflected in the distribution of both malignancy and lung nodule visual attributes (LNVAs), as detailed in Appendix .3 (Tables A.2 and A.3). In addition, several attributes present strong class imbalance, with certain classes overwhelmingly dominating the distribution (Table A.4). A similar pattern is also observed in the LUNA25, where malignancy labels are also highly skewed towards the benign class (Table A.5).

Together, these characteristics highlight two complementary challenges: noisy supervision arising from inter-observer disagreement in LIDC-IDRI, and biased learning caused by imbalanced class distributions across both datasets.

3.2.4 Data Splitting Strategy

The scale used by the radiologists to annotate malignancy in the LIDC-IDRI dataset goes from 1 to 5, from 'highly unlikely' to 'highly suspicious' nodules (3.2.1). In this work, a single label is derived per nodule by aggregating the individual annotations using the median rather than the mean, as the median is more robust to inter-observer variability. To obtain a binary classification setting, nodules with median malignancy scores ≤ 2 are considered benign, and those with scores ≥ 4 are considered malignant. Nodules with a median score of 3 represent ambiguous cases, reflecting substantial disagreement or uncertainty among radiologists so, even if they constitute a good portion of the data, as shown in Appendix .3 (Table A.2), these samples were excluded from stratification, like in prior work [22], and assigned exclusively to the training set, not being used for malignancy evaluation, but remaining available during training for LNVA supervision.

For the LUNA25 dataset, training and validation splits were obtained using stratified sampling based on malignancy labels. Within the semi-supervised framework, the unlabeled samples used for consistency regularization were drawn exclusively from the training partition, ensuring that no validation data is used during training. This design guarantees that both supervised and semi-supervised components operate strictly within the training data, preventing information leakage and ensuring a reliable evaluation protocol.

3.3 Multi-Task Learning Setup

The supervised learning framework is based on two differently labeled, but complementary datasets. The LIDC-IDRI dataset provides full supervision for both malignancy and lung nodule visual attributes, while LUNA25 contributes supervision only for malignancy classification. Then, within the proposed semi-supervised learning framework, the latter is additionally treated as unlabeled data. Attribute predictions obtained from these samples are used in a consistency regularization strategy, enabling the transfer of attribute-related knowledge learned from LIDC-IDRI to LUNA25.

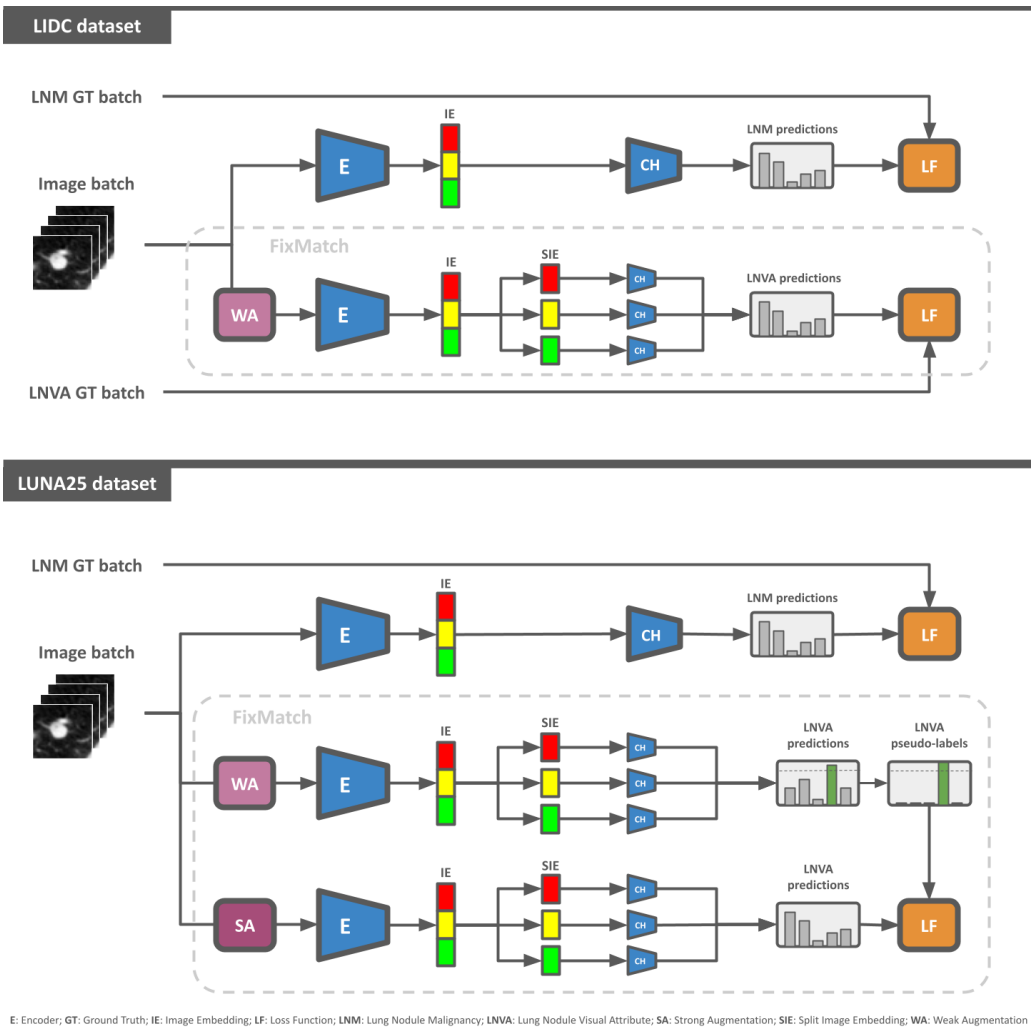


FIGURE 3.3: **Overview of the proposed pipeline across different datasets: shared backbone, multi-task heads and the semi-supervised learning branch.** *Top (LIDC dataset):* Supervised training for malignancy and the supervised part of fixmatch, using lung nodule visual attributes. *Bottom (LUNA25 dataset):* Supervised training for malignancy and the semi-supervised learning framework utilizing fixMatch. **Abbreviations: E**: Encoder; **GT**: Ground Truth; **IE**: Image Embedding; **LF**: Loss Function; **LNM**: Lung nodule malignancy; **LNVA**: Lung nodule visual attribute; **SA**: Strong Augmentation; **SIE**: Split Image Embedding; **WA**: Weak Augmentation.

All tasks are learned jointly through a shared backbone, which couples malignancy prediction and attribute learning within a common feature space. In this setting, attribute supervision from LIDC-IDRI guides the model towards learning more descriptive representations, while LUNA25 contributes to improving malignancy prediction under a different data distribution. The semi-supervised branch further constrains the attribute predictions by enforcing consistency on unlabeled samples.

The semi-supervised strategy was inspired by FixMatch [14], where for each unlabeled

sample x_i, two augmented versions are generated: a weakly augmented view x_i^w and a strongly augmented view x_i^s, following the FixMatch paradigm (see Appendix A.1). The strong augmentation is used to enforce consistency, while the weak augmentation is used to produce a pseudo-label defined as:

$$\hat{y}_i = \arg\max p(y \mid x_i^w) \tag{3.1}$$

If the prediction with maximum confidence is higher than a threshold τ, the pseudo-label is used as a supervision signal for the strongly augmented sample. Considering $H(\cdot)$ as the the cross-entropy loss and B_u as the batch of unlabeled samples, the resulting loss is then defined as:

$$\mathcal{L}_{SSL} = \frac{1}{|B_u|} \sum_i \mathbf{1} \left[\max p(y \mid x_i^w) \geq \tau \right] \cdot H(\hat{y}_i, f(x_i^s)) \tag{3.2}$$

While FixMatch provides a principled baseline for consistency regularization, its reliance on hard confidence thresholding introduces a structural limitation when applied to multi-class ordinal prediction tasks such as LNVA attribute learning. In this setting, the predicted probability mass is distributed across five ordinal categories (see Appendix .1 Table A.1), possibly making it rare for any single class to exceed a high confidence threshold. This motivates the adoption of UniMatch [25], a dual-stream consistency framework that addresses this limitation by generating two independently strongly-augmented views of each unlabeled sample, rather than relying exclusively on a weakly-augmented pseudo-label as the supervision signal:

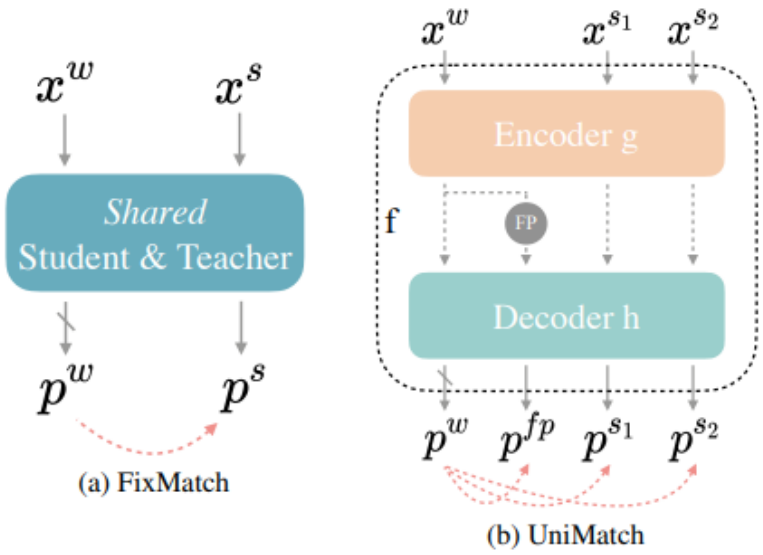


FIGURE 3.4: FixMatch vs UniMatch framework. Adapted from [25].

The corresponding loss function is defined as:

$$\mathcal{L}_{SSL}^{UNI} = \frac{1}{3|B_u|} \sum_i m_i \left[H(\hat{y}_i, f(x_i^{s1})) + H(\hat{y}_i, f(x_i^{s2})) + H(\hat{y}_i, f(x_i^{fd})) \right] \tag{3.3}$$

where m_i is the mask for the i-th sample, x_i^{s1} and x_i^{s2} are its two independent strong augmentations, and x_i^{fd} is the version with feature dropout.

3.4 Proposed Architecture

A convolutional encoder $g(\cdot)$ maps each input patch x into a latent representation $z = g(x)$. This shared representation is then used by multiple task-specific heads: a malignancy classification head predicting y_m, multiple LNVA attribute heads predicting each attribute a_k and a semi-supervised branch enforcing consistency over unlabeled samples. This design ensures that all tasks operate in a shared feature space, encouraging the learning of generalizable radiological representations.

Given the variability in lesion appearance, scale and texture in lung nodule analysis, the choice of convolutional backbone plays a critical role in representation learning. In this work, different backbone architectures were investigated, including ResNet, DenseNet and Res2Net, to assess how different inductive biases affect both malignancy classification and auxiliary LNVA prediction:

- **ResNet**: provides a residual learning framework that enables stable optimization of deep networks through skip connections. This design mitigates vanishing gradients and facilitates the training of deeper architectures.
- **DenseNet**: promotes feature reuse through dense connectivity, improving the propagation of fine-grained information across layers. This is particularly beneficial for capturing subtle radiological patterns in medical imaging.
- **Res2Net**: extends the residual paradigm by introducing multi-scale feature extraction within each block, enabling the capture of features at different receptive fields simultaneously.

In the supervised setting, ResNet-10, ResNet-18 and ResNet-34 were used to evaluate best suited model capacity. DenseNet and Res2Net are further compared against the ResNet baseline, in the multi-task phase, to analyze the impact of different architectural biases on both malignancy and LNVA learning.

3.5 Training Objective

The training objective is defined as a combination of supervised and consistency-based losses, reflecting the heterogeneous nature of the available annotations and the chosen multi-task structure of the model. This combination focuses on two main components: malignancy prediction and visual attribute learning. The former is defined as the sum of the fully supervised malignancy losses computed using LIDC-IDRI samples ($\mathcal{L}_{LIDC}$) and LUNA25 samples ($\mathcal{L}_{LUNA}$), while the latter corresponds to a combination of LNVA attribute supervision, computed only on LIDC-IDRI samples ($\mathcal{L}_S$), and a consistency regularization loss applied to LNVA predictions for LUNA25 samples ($\mathcal{L}_{SSL}$).

It is important to note that the visual attribute losses are defined by calculating the mean over multiple attribute-specific prediction heads, and that each loss component is associated with an individual weighting factor (λ_*) to balance their relative contributions during training.

$$\mathcal{L} = \underbrace{\lambda_{LIDC}\mathcal{L}_{LIDC} + \lambda_{LUNA}\mathcal{L}_{LUNA}}_{\mathcal{L}_{\text{malignancy}}} + \underbrace{\lambda_S\mathcal{L}_S}_{\mathcal{L}_{\text{multi-task}}} + \underbrace{\lambda_{SSL}\mathcal{L}_{SSL}}_{\mathcal{L}_{\text{SSL}}} \tag{3.4}$$

Each component plays a different role in the learning process. The LIDC-IDRI branch provides fine-grained supervision for both malignancy and visual attributes, enabling

the model to learn detailed radiological patterns. In contrast, the LUNA25 branch introduces additional malignancy supervision under a different data distribution and further contributes through the consistency loss, which regularizes the attribute heads. The fixed weighting factors ensure that all sources contribute without dominating the optimization process.

3.6 Experimental Design Strategy

To systematically evaluate the contribution of each component in the proposed framework, the experimental protocol follows a progressive design, where each stage introduces a single additional source of supervision or regularization. This allows the individual contribution of LNVA learning and semi-supervised consistency constraints to be isolated and analyzed independently.

Phase 1 — Supervised Foundation

The first phase of the experimental design serves as a controlled reference point that characterizes the behavior of standard supervised learning and provides the foundation upon which more complex multi-task and semi-supervised strategies are later introduced.

In medical imaging problems such as lung nodule classification, performance is highly sensitive to both the model capacity and training stability, resulting in the evaluation of different backbone architectures alongside learning rates to understand the trade-off between representational power and generalization. Similarly, normalization strategies are compared to assess their robustness when relying on either batch-dependent statistics or channel-wise normalization under constrained and heterogeneous training conditions.

Beyond architectural considerations, the choice of loss function is also treated as a key design variable, as class imbalance is a known and the biggest challenge in LUNA25. Different loss formulations are therefore explored to determine whether alternative optimization objectives can improve robustness to skewed label distributions.

Phase 2 — Multi-Task Learning

The second phase of the experimental design investigates the contribution of auxiliary lung nodule visual attributes supervision to malignancy classification introducing a multi-task learning formulation where shared representations are jointly optimized for both, malignancy prediction and attribute estimation.

The primary objective of this phase is to evaluate whether incorporating LNVA supervision leads to improved generalization in malignancy classification. To do this, we vary the strength of the visual attributes loss to asses its impact on representation learning and compare standard loss formulations with disagreement-aware weighting schemes aiming for a more robust feature learning. This allows us to understand not only the impact of LNVA's learning on the malignancy, but also if the model is learning this attributes successfully in the first place.

To further validate the robustness of the multi-task formulation, we perform a systematic backbone comparison under a fixed multi-task supervised setting. Specifically, we evaluate ResNet, DenseNet, and Res2Net variants while keeping all other training components constant. The objective is to isolate the effect of the feature extractor on the prediction of malignancy and lung nodule visual attributes, and to determine whether certain architectural biases are more suited to capturing the shared representation required for both tasks.

Phase 3 — Semi-Supervised Learning

This phase extends the multi-task framework by incorporating semi-supervised learning using FixMatch based consistency regularization. The objective is to investigate whether unlabeled data (LUNA25) can further improve lung nodule visual attribute learning and potentially malignancy classification as well, without degrading the supervised signal obtained from LIDC-IDRI.

The semi-supervised formulation builds upon the best-performing multi-task configuration selected in Phase 2, where unlabeled samples are incorporated through pseudo-labeling based on high-confidence predictions. Consistency is enforced between weakly and strongly augmented views of the same input, encouraging the model to learn more

stable representations across different data distributions.

To evaluate the sensitivity of the framework, we vary the FixMatch confidence threshold and the weighting of the consistency loss, allowing us to analyze the impact of pseudo-label quality and contribution strength during training.

To address the structural limitations identified in the static configuration, specifically low pseudo-label utilization and early training instability, this phase introduces a joint threshold and weight scheduling strategy. This advanced setup incorporates four concurrent modifications: (i) *threshold scheduling*, to maximize data coverage early in training; (ii) *an SSL weight ramp-up*, to prevent early representation disruption; (iii) *a reduced LNVA loss weight* to free representational capacity; and (iv) *selective attribute supervision*, restricting both supervised and consistency targets strictly to the most predictive visual features.

Finally, recognizing that hard confidence thresholding struggles with multi-class ordinal targets where probability mass is widely distributed, the phase culminates in the transition to UniMatch, a dual-stream consistency framework that generates two independently strongly-augmented views alongside a feature-dropout perturbation in the latent space, reducing reliance on hard thresholding as the sole mechanism for pseudo-label selection.

Phase 4 — Interpretability Analysis and Attribute Contribution

The final phase of the experimental design focuses on model interpretability, aiming to understand how the learned representations relate to clinically meaningful radiological patterns and how they contribute to the malignancy prediction. Here, we investigate interpretability from two complementary perspectives: First, we analyze the spatial behavior of the model using Gradient-weighted Class Activation Mapping (Grad-CAM), applied independently to the malignancy prediction head and to each LNVA prediction head. This allows us to examine whether the model focuses on clinically relevant regions when predicting both malignancy and specific visual attributes. Second, we study the relationship between LNVA predictions and the final malignancy decision in a quantitative manner. The goal is to assess whether the learned attribute representations carry predictive information about malignancy and to what extent each attribute contributes to the final decision.

Chapter 4

Experimental Setup

This chapter describes the practical conditions under which all experiments were conducted. It covers the computational environment and software stack, the data preprocessing pipeline, the training configuration, and the loss functions evaluated across tasks. It further details the augmentation strategies used in the semi-supervised framework, the experimental configuration that instantiates the staged design introduced in Chapter 3, the evaluation protocol applied throughout, and closes with a description of the logistic regression analysis used to quantify the contribution of each visual attribute to malignancy prediction.

4.1 Computational Environment

All experiments were conducted on a cluster managed via SLURM (Simple Linux Utility for Resource Management), and the jobs themselves were executed on NVIDIA GPUs ($\geq$ 24GB VRAM), such as A5000 and A6000 devices and had 12 CPU cores allocated per task. Mixed-precision training (FP16) was enabled through PyTorch Lightning by setting "precision=16-mixed", significantly reducing memory consumption while maintaining numerical stability during optimization.

4.2 Software Stack

The experimental framework was implemented using PyTorch 2.2.2 with CUDA 12.1 support, ensuring compatibility with modern GPU-accelerated computation. Training and evaluation pipelines were built using PyTorch Lightning, which provided structured

training loops. To enable definition of models, datasets and training parameters via command-line overrides, `Hydra` [*] was used to manage the experiments configurations. The tracking was done using `MLflow` [†] and additional supporting libraries include NumPy and Scikit-learn, used for numerical processing and evaluation metrics.

4.3 Data Preprocessing

Input data consists of two-dimensional lung nodule patches extracted from CT scans, stored as NumPy arrays. Each sample is reshaped into a single-channel tensor of dimensions $(1, H, W)$ and resized to 64×64 pixels to ensure consistent spatial resolution across datasets. Intensity values are normalized using fixed mean and standard deviation normalization (mean = 0.5, std = 0.5), ensuring consistent scaling across all inputs.

Annotations are obtained from multiple radiologists and aggregated using the median value and the LNVAs are encoded in an ordinal format. Subsequently, the LNVAs are encoded in an ordinal format and shifted to a zero-based representation for compatibility with the classification heads. As discussed before, cases with a median malignancy score equal to 3 are treated as ambiguous and excluded from the malignancy loss computation, while remaining available for LNVA supervision. Although a subset of these attributes can be treated as ordinal structures, most are fundamentally categorical in nature and so, for consistency and simplicity of the modelling framework, all attributes are treated as such.

4.4 Training Setup

To ensure consistency across experiments, we started with a predefined set of training parameters as summarized in Table 4.1.

[*] Hydra library website: https://hydra.cc/docs/intro/
[†] MLflow library website: https://mlflow.org/

TABLE 4.1: Training hyperparameters used across all experiments

Component	**Configuration**
Optimizer	Adam
Learning rate	$\{1, 3, 5\} \times 10^{-4}$
Weight decay	0.01
LIDC batch size	32
LUNA batch size	128
Max epochs	60
Scheduler	ReduceLROnPlateau
Scheduler patience	6
Scheduler factor	0.5
Min learning rate	1×10^{-7}
Seeds	{42, 124, 168, 336, 672}
Model selection metric	Validation AUROC

This configuration is kept fixed across all experiments unless explicitly stated otherwise, ensuring that performance differences arise from architectural or methodological changes rather than training instability or hyperparameter variation.

4.5 Loss Functions

Throughout this section, all loss functions are denoted using the $\mathcal{L}$ notation, where task-specific losses are indexed by their respective components (e.g., malignancy, LNVA, SSL).

4.5.1 Cross-Entropy-based Losses

Cross-Entropy is used as the baseline optimization objective. Binary cross-entropy is applied to malignancy prediction, while categorical cross-entropy is used for multi-class LNVA attribute prediction, where C denotes the number of classes, y_c the one-hot encoded ground-truth label and $\hat{y}_c$ the predicted probability for class c.

$$\underbrace{\mathcal{L}_{\text{CE}} = -\left[y \log(\hat{y}) + (1-y)\log(1-\hat{y})\right]}_{\text{Binary CE (malignancy classification)}} \qquad \underbrace{\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c)}_{\text{Categorical CE (LNVA prediction)}}$$

To mitigate class imbalance, Weighted Cross-Entropy is additionally explored, where each class is assigned a weighting factor inversely proportional to its frequency in the training set. The class weights are defined as $w_c = \frac{N}{C \cdot n_c}$, where N is the total number of samples, C is the number of classes and n_c is the number of samples belonging to class c. The weighted loss is then applied to the standard cross-entropy formulation:

$$\mathcal{L}_{\text{WCE}} = -\sum_{c=1}^{C} w_c \, y_c \log(\hat{y}_c)$$

This weighting strategy increases the contribution of underrepresented classes during optimization, reducing the bias toward majority-class predictions.

4.5.2 Alternative Robust Losses

To further improve robustness under class imbalance and noisy annotations, as seen in Appendix .3 (Tables A.4 and A.3), additional loss functions are explored beyond standard cross-entropy. These include Focal Loss, which dynamically down-weights easy samples, and Cross-entropy with Dice, which combines region-overlap and classification objectives to improve learning stability in imbalanced settings. The latter introduces a modulating factor to the standard cross-entropy, reducing the contribution of well-classified samples and focusing training on harder examples. Given γ as a focusing parameter that controls the strength of down-weighting easy samples, it is defined as:

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

Cross-Entropy with Dice regularization combines the standard cross-entropy loss with the Dice coefficient, encouraging both correct classification and better overlap between predicted and ground-truth distributions. Given a λ_D that controls the contribution of the Dice term, it is defined as:

$$\mathcal{L}_{\text{CE-Dice}} = \mathcal{L}_{\text{CE}} + \lambda_D \left(1 - \frac{2\sum y\hat{y}}{\sum y + \sum \hat{y}}\right)$$

Disagreement-aware LNVA Losses

To explicitly account for inter-observer variability in LNVA annotations, we created a disagreement-aware weighting loss based on normalized entropy. Let H_a denote the normalized entropy of attribute a, computed over the empirical distribution of radiologist annotations. This defines an attribute-level reliability weight:

$$w_a = 1 - (H_a + \epsilon), \qquad \text{with } \epsilon = 10^{-6} \text{ to avoid degenerate cases where } H_a \approx 1.$$

Using this weighting, we define a general disagreement-aware cross-entropy over all attributes a and classes c, where $w_{a,c}$ denotes the class-specific weighting factor:

$$\mathcal{L} = -\sum_{a=1}^{A} \underbrace{w_a \sum_{c=1}^{C_a} y_{a,c} \log(\hat{y}_{a,c})}_{\text{DCE}} \quad \text{and} \quad \mathcal{L} = -\sum_{a=1}^{A} \underbrace{w_a \sum_{c=1}^{C_a} w_{a,c}\, y_{a,c} \log(\hat{y}_{a,c})}_{\text{DWCE}}$$

These formulations explicitly couple inter-observer disagreement and class imbalance, ensuring that both annotation uncertainty and data frequency are accounted for during optimization.

Loss Assignment Across Tasks

The application of each loss function depends on the prediction task. Let $\mathcal{L}_{\text{mal}}$ denote the malignancy loss space and $\mathcal{L}_{\text{lnva}}$ the LNVA loss space. The mapping is defined as:

$$\mathcal{L}_{\text{mal}} \in \{\text{CE}, \text{WCE}, \text{Focal}, \text{CE-Dice}\}$$
$$\mathcal{L}_{\text{lnva}} \in \{\text{CE}, \text{WCE}, \text{DCE}, \text{DWCE}\}$$

This separation reflects the different nature of the two tasks: malignancy prediction relies on fully supervised labels and therefore focuses on robustness under class imbalance,

while LNVA prediction additionally incorporates inter-observer disagreement, motivating the use of disagreement-aware formulations.

4.6 Semi-Supervised Augmentation Strategies

This section describes the augmentation pipelines used across the two semi-supervised frameworks evaluated in this work: FixMatch and UniMatch. Both share a common weak augmentation policy used to generate pseudo-labels, but differ in how they construct the consistency targets.

4.6.1 Augmentations

FixMatch relies on two distinct augmentation policies to generate pseudo-label supervision from unlabeled data. In this work, both weak and strong augmentation pipelines are designed specifically for medical imaging, balancing structural preservation with controlled perturbations. The detailed augmentation operators for each pipeline are summarized in Table 4.3, and a visual comparison is shown in Figure 4.1.

TABLE 4.2: Weak and strong augmentation policies used in FixMatch training.

Operation	**Weak Augmentation**	**Strong Augmentation**
Resize	✓	✓
Horizontal Flip	$p = 0.5$	$p = 0.5$
Vertical Flip	$p = 0.25$	$p = 0.25$
Rotation (90° steps)	RandomChoice	RandomChoice
Gaussian Blur	kernel=3	kernel=5, $\sigma \in [0.5, 2.0]$
Sharpness Adjustment	–	$p = 0.5$, factor=2
Random Erasing	–	$p = 0.5$, scale=$(0.02, 0.2)$
Normalization	$\mu = 0.5$, $\sigma = 0.5$	$\mu = 0.5$, $\sigma = 0.5$

TABLE 4.3: Weak and strong augmentation policies used in FixMatch training.

Although Random Erasing may appear aggressive for medical imaging, it is applied with a strictly constrained spatial range (2%–20% of the image area), preserving global lesion structure and boundary morphology.

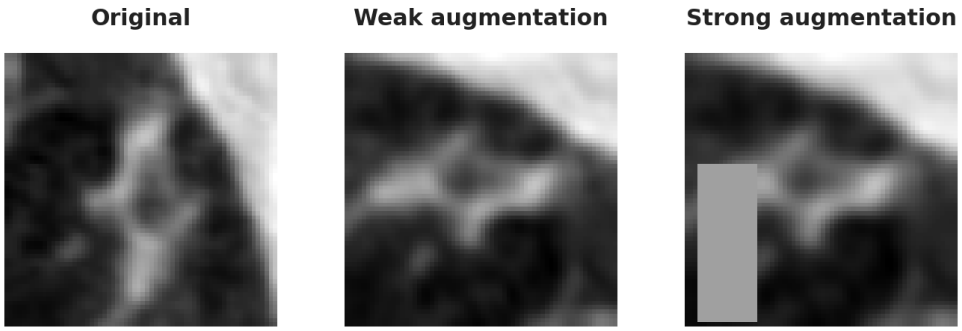


FIGURE 4.1: Comparison between weak and strong augmentation pipelines used in FixMatch. From left to right: original patch, weak augmentation, strong augmentation.

UniMatch extends the FixMatch augmentation strategy by generating three consistency targets per unlabeled sample instead of one: two independently sampled strong augmentations and a feature-space perturbation obtained via dropout applied directly in the encoder representation. The weak augmentation policy remains identical to FixMatch and is used solely to produce the pseudo-label. The two strong augmentations use the same operator set as the FixMatch strong pipeline (Table 4.3), but are sampled independently, meaning that different random parameters are drawn for each view. This ensures that the two views present genuinely distinct perturbations of the same input, increasing the diversity of the consistency constraint without requiring additional augmentation operators.

4.6.2 Comparative Overview

To make the structural difference between the two frameworks explicit, Fixmatch 1 and Unimatch 2 present their respective consistency regularization procedures side by side. Both methods share the same pseudo-label generation mechanism based on weak augmentation and confidence thresholding; the key distinction lies in how the consistency loss is constructed from the unlabeled sample.

Algorithm 1: FixMatch

Require: sample x, threshold τ, model f

$x^w \leftarrow \texttt{augment\_weak}(x)$
$x^s \leftarrow \texttt{augment\_strong}(x)$
$\hat{y} \leftarrow \arg\max\ p(y \mid x^w)$ ▷ pseudo-label
$m \leftarrow \left[\max p(y \mid x^w) \geq \tau\right]$ ▷ mask
$\mathcal{L}_{\text{SSL}} \leftarrow m \cdot H(\hat{y}, f(x^s))$

Algorithm 2: UniMatch

Require: sample x, threshold τ, model f

$x^w \leftarrow \texttt{augment\_weak}(x)$
$x^{s_1} \leftarrow \texttt{augment\_strong}(x)$
$x^{s_2} \leftarrow \texttt{augment\_strong}(x)$
$x^{fd} \leftarrow \texttt{feature\_dropout}(x^w)$
$\hat{y} \leftarrow \arg\max\ p(y \mid x^w)$ ▷ pseudo-label
$m \left[\max p(y \mid x^w) \geq \tau\right]$ ▷ mask
$\mathcal{L}_{\text{SSL}} \leftarrow \frac{m}{3} \sum_v H\left(\hat{y}, f(x^v)\right)$ para $v \in \{s_1, s_2, fd\}$

Figure 4.2: Pseudocode comparison for each unlabeled sample processed by the FixMatch and UniMatch frameworks.

The key differences between the two approaches are threefold. First, UniMatch introduces a second independent strong augmentation, which diversifies the perturbation space covered by the consistency constraint without requiring any additional labeled data or architectural changes. Second, the feature dropout view provides a perturbation in the latent space, decoupling the consistency signal from purely pixel-level variation and making it more robust to input distribution shift. Third, by averaging the loss over three targets rather than one, UniMatch reduces the variance of the gradient estimate from unlabeled samples, which can stabilize training when pseudo-label quality is heterogeneous.

4.7 Experimental Configuration

All experiments follow the staged evaluation strategy described in Chapter 3, Section 3.6. The Phase 1 — Supervised Foundation evaluates multiple backbonearchitectures and optimization settings. ResNet-based architectures are compared under learning rates $\{1 \times 10^{-4}; 3 \times 10^{-4}; 5 \times 10^{-4}\}$, different normalization strategies and alternative malignancy loss formulations. Phase 2 — Multi-Task Learning extends the selected supervised configuration with LNVA prediction heads. Different attribute-loss weighting schemes, including disagreement-aware formulations, are evaluated while keeping the malignancy objective unchanged.

The Phase 3 — Semi-Supervised Learning builds upon the best-performing multi-task

configuration and proceeds in three stages. First, a grid search over static FixMatch hyperparameters is performed: confidence thresholds $\tau \in \{0.85, 0.90, 0.95\}$ and consistency weights $\alpha \in \{0.5, 1.0, 2.0\}$. Second, motivated by the low pseudo-label utilization observed in the static configuration, a joint threshold and weight scheduling strategy is introduced, combining linear threshold annealing ($\tau_{\text{start}} = 0.60 \rightarrow \tau_{\text{end}} = 0.85$), an SSL weight ramp-up over the first 30 epochs, a reduced LNVA loss weight $\alpha_{\text{LNVA}} \in \{1.0, 1.5\}$, and selective attribute supervision restricted to the five most predictive LNVAs. A sweep over $\alpha_{\text{FixMatch}} \in \{0.5, 1.0, 1.5\}$ is conducted under this scheduled configuration. Third, UniMatch is evaluated as an alternative to hard-threshold pseudo-labeling, replacing the single strongly-augmented consistency target with a dual-stream formulation; feature dropout rates $\in \{0.0, 0.1, 0.3, 0.5\}$ are swept under the best scheduling configuration.

Finally, Phase 4 — Interpretability Analysis and Attribute Contribution is conducted exclusively on the canonical FixMatch semi-supervised model to ensure that visual explanations correspond to a stable and fully validated representation. Grad-CAM visualizations are extracted from both the malignancy and LNVA heads, and attribute contribution is assessed through post-hoc logistic regression analysis of the learned representations. A visual representation of the experimental configuration can be seen in Figure 4.3

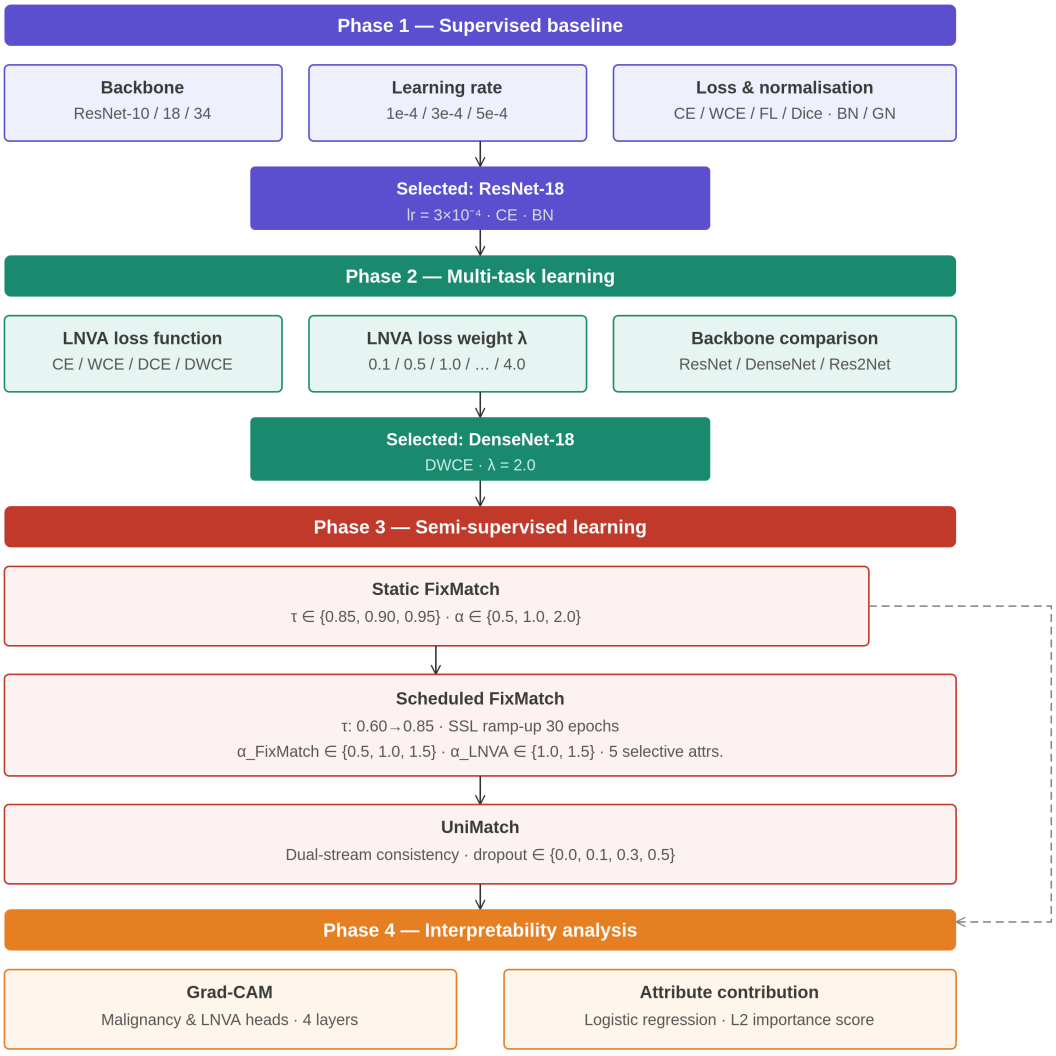


FIGURE 4.3: Overview of the staged experimental design, progressing from supervised baseline selection through multi-task learning and semi-supervised extensions to interpretability analysis. The dashed arrow indicates the model used for the interpretability phase.

4.8 Evaluation Protocol

All experiments are evaluated under identical data splits and training conditions unless otherwise specified. Performance is reported using AUROC, f_1 score and Accuracy. Initially, in the single-task supervised learning these metrics are computed separately for

LIDC-IDRI and LUNA25, due to their differing label noise and class imbalance characteristics and then jointly calculated in the next phases. To ensure robustness, all experiments are repeated using five fixed random seeds $\{42, 124, 168, 336, 672\}$, and results are reported as mean and standard deviation between those.

4.9 Contribution Analysis via Logistic Regression

To quantify the predictive contribution of each LNVA attribute to malignancy classification, we trained a set of linear logistic regression models using LNVA prediction outputs as input features. For each attribute a, let $\mathbf{x}_a \in \mathbb{R}^C$ denote the predicted probability distribution over its C classes, extracted from the trained multi-task model. These probability vectors are used as features to predict the binary malignancy label $y \in \{0, 1\}$.

Each logistic regression model is trained independently per attribute using a held-out validation split. Features are standardized using z-score normalization prior to training, and an L2-regularized logistic regression model (C = 1.0) is fitted using the `scikit-learn` library* implementation. The importance score for each attribute is computed as the L2 norm of the learned coefficient vector:

$$S_a = \|\mathbf{w}_a\|_2$$

where $\mathbf{w}_a$ denotes the learned logistic regression weights for attribute a.

*Library website: https://scikit-learn.org/stable/

Chapter 5

Results and Discussion

The experimental evaluation follows the staged design described in Chapter 4. Each section builds directly on the configuration selected in the previous one, allowing the contribution of each component to be assessed in isolation before the next layer of complexity is introduced.

5.1 Supervised Baseline

This section reports the performance of a supervised malignancy classification baseline, trained jointly on LIDC-IDRI and LUNA25 under varying backbone architectures and learning rates. All results are reported as mean $\pm$ standard deviation over five random seeds, using AUROC and accuracy as primary evaluation metrics.

Architecture and Learning Rate Selection

The performance of ResNet-10, ResNet-18 and ResNet-34 across three learning rates is reported in Table 5.1. For the entirety of this chapter, metrics are going to be reported separately per dataset for clarity, although the model is trained jointly and the total loss is computed as the sum of both datasets. This separation is important because one dataset is specially noisy, while the other highly imbalanced, making it useful to assess how the model behaves on each dataset individually.

TABLE 5.1: Supervised baseline results across architectures and learning rates.

Architecture	LR	Dataset	Accuracy	AUROC
ResNet-10	0.0001	LIDC	0.8743 ± 0.0267	0.9011 ± 0.0058
		LUNA	0.7178 ± 0.0734	0.8544 ± 0.0101
	0.0003	LIDC	0.8934 ± 0.0079	0.8947 ± 0.0044
		LUNA	0.7750 ± 0.0565	0.8533 ± 0.0051
	0.0005	LIDC	**0.9099** ± 0.0053	0.8939 ± 0.0077
		LUNA	0.8058 ± 0.0170	0.8497 ± 0.0060
ResNet-18	0.0001	LIDC	0.8987 ± 0.0060	0.8911 ± 0.0041
		LUNA	0.7655 ± 0.0427	0.8533 ± 0.0073
	0.0003	LIDC	0.9026 ± 0.0191	0.8993 ± 0.0068
		LUNA	0.7624 ± 0.0499	**0.8647** ± 0.0085
	0.0005	LIDC	**0.9099** ± 0.0053	0.8941 ± 0.0036
		LUNA	0.7786 ± 0.0237	0.8536 ± 0.0070
ResNet-34	0.0001	LIDC	0.9026 ± 0.0107	0.9004 ± 0.0085
		LUNA	**0.8075** ± 0.0536	0.8557 ± 0.0055
	0.0003	LIDC	0.9013 ± 0.0139	**0.9047** ± 0.0029
		LUNA	0.8064 ± 0.0392	0.8592 ± 0.0100
	0.0005	LIDC	0.8536 ± 0.0221	0.8520 ± 0.0115
		LUNA	0.8015 ± 0.0254	0.7761 ± 0.0192

* *Note: This runs are using batch normalization and cross entropy*

Overall, all three architectures achieve comparable performance on LIDC-IDRI, having AUROC values consistently above 0.89, with the more notable, but small differences emerge when evaluating generalization on LUNA25. ResNet-34 exhibits a marked degradation at higher learning rates, with AUROC dropping to 0.7761 on LUNA25. The other two architectures demonstrate more consistent performance across datasets with ResNet-18 achieving a better trade-off between strong performance and stable generalization. Regarding the learning rates, $lr = 1 \times 10^{-4}$ yields marginally higher AUROC on LIDC-IDRI,

but consistently underperforming on LUNA25, and because $lr = 5 \times 10^{-4}$ tends to destabilize training for deeper networks, $lr = 3 \times 10^{-4}$ was chosen as the most stable overall. Given its favorable trade-off between predictive performance, robustness and computational efficiency, ResNet-18 with $lr = 3 \times 10^{-4}$ was then selected as the backbone configuration for all subsequent experiments.

Normalization Strategy

Table 5.2 compares Batch and Group Normalization under identical supervised training conditions using ResNet-18 with $lr = 3 \times 10^{-4}$. There, Batch Normalization consistently outperforms the Group approach across both datasets, on the evaluation metrics and standard deviation.

TABLE 5.2: Effect of normalization strategy on supervised ResNet-18 performance

Normalisation	Dataset	Accuracy	AUROC
Batch Norm	LIDC	**0.9132** $\pm$0.0068	**0.8832** $\pm$0.0069
	LUNA	**0.7596** $\pm$0.0269	**0.7835** $\pm$0.0083
Group Norm	LIDC	0.8954 ± 0.0176	0.8723 ± 0.0163
	LUNA	0.7411 ± 0.0612	0.7774 ± 0.0083

These results suggest that the batch sizes employed during training are sufficient for Batch Normalization to estimate stable statistics. Consequently, Batch Normalization was adopted as the default normalization strategy.

Malignancy loss formulation

Table 5.3 summarizes the impact of different malignancy loss formulations under the fully supervised setting using the ResNet-18 backbone. We are testing various losses with the main objective of dealing with the imbalance of the LUNA25 dataset, and therefore, it makes sense to also look at the f_1 score for that dataset. This metric is stable across all losses except for Weighted CE, possibly due to over penalization.

TABLE 5.3: Effect of malignancy loss functions on supervised ResNet-18 performance.

Loss Function	Dataset	Accuracy	AUROC	f_1 score
CE (4.5.1)	LIDC	0.9046 ± 0.0042	**0.8756** ± 0.0229	–
	LUNA	0.7606 ± 0.0450	0.7783 ± 0.0104	0.6152 ± 0.0300
Weighted CE (4.5.1)	LIDC	0.8829 ± 0.0447	0.8739 ± 0.0237	–
	LUNA	0.6964 ± 0.0759	0.7739 ± 0.0185	0.5737 ± 0.0522
Focal Loss (4.5.2)	LIDC	0.8987 ± 0.0032	0.8628 ± 0.0115	–
	LUNA	**0.7698** ± 0.0374	**0.7859** ± 0.0097	**0.6238** ± 0.0284
BCE-Dice (4.5.2)	LIDC	**0.9086** ± 0.0109	0.8721 ± 0.0042	–
	LUNA	0.7638 ± 0.0262	0.7753 ± 0.0039	0.6155 ± 0.0174

Overall, the evaluated loss functions achieve broadly comparable performance, suggesting that the supervised baseline is relatively robust to the choice of optimization objective. There is an argument to me made that CE and Focal are superior than the others, with the former looking like the better option. But being the differences so small, and because the LNVA and semi-supervised losses are going to be CE based, we will be using the CE loss moving forward, avoiding loss scales imbalance.

5.2 Multi-Task Learning

Building on the supervised baseline, this section investigates the impact of auxiliary learning signals on malignancy classification performance through multi-task learning. The motivation for this approach lies in understanding the trade-off between improved interpretability, by encouraging the model to learn clinically meaningful representations, and predictive performance in malignancy classification. As in Section 5.1, all results are reported as mean $\pm$ standard deviation over five random seeds. The experiments uses the configuration selected in the previous section:

TABLE 5.4: Selected configuration for multi-task learning experiments.

Component	**Setting**
Backbone	ResNet-18
Learning Rate	3×10^{-4}
Normalization	Batch Normalization
Loss Function	Cross-Entropy
Loss Weights	LUNA 0.5 — LIDC 0.5

Loss Formulation and Task Weighting Strategy

Table 5.5 reports the performance of different LNVA loss formulations under the multi-task learning setting, evaluated across multiple task weighting values (λ). The results are no longer divided per dataset, as we are now looking at general performance. And since the macro-f_1 for malignancy on the LUNA25 dataset remained stable in the previous section, it is omitted from subsequent tables.

TABLE 5.5: Average malignancy classification performance across LIDC-IDRI and LUNA25 datasets for different LNVA multitask loss functions and weighting values (λ).

LNVA Loss	λ	**Accuracy**	**AUROC**
CE (4.5.1)	0.1	0.8496 ± 0.0161	$\mathbf{0.8305} \pm 0.0122$
	0.5	0.8503 ± 0.0277	0.8222 ± 0.0184
	2.0	$\mathbf{0.8619} \pm 0.0229$	0.8029 ± 0.0282
DisagreementCE (4.5.2)	0.1	0.8120 ± 0.0334	$\mathbf{0.8297} \pm 0.0100$
	0.5	$\mathbf{0.8625} \pm 0.0197$	0.8144 ± 0.0126
	2.0	0.8504 ± 0.0258	0.8156 ± 0.0207
WeightedCE (4.5.1)	0.1	0.8106 ± 0.0098	$\mathbf{0.8345} \pm 0.0070$
	0.5	0.8444 ± 0.0219	0.8281 ± 0.0292
	2.0	$\mathbf{0.8626} \pm 0.0167$	0.7894 ± 0.0191
DisagreementWCE (4.5.2)	0.1	0.8252 ± 0.0356	$\mathbf{0.8301} \pm 0.0154$
	0.5	0.8456 ± 0.0336	0.8156 ± 0.0249
	2.0	$\mathbf{0.8533} \pm 0.0167$	0.8159 ± 0.0171
(Supervised Baseline)		0.8326 ± 0.0246	0.8225 ± 0.0166

Multi-task learning improves or maintains malignancy classification performance when compared to the fully supervised baseline. In fact, at lower weighting values ($\lambda = 0.1$), AUROC consistently exceeds the supervised reference across all four loss formulations, suggesting that a modest degree of LNVA supervision provides useful inductive bias for malignancy prediction without disrupting the primary objective. However, as λ increases, a clear trade-off emerges: accuracy improves at the cost of AUROC. This suggests that stronger emphasis on attribute supervision encourages more confident class predictions, but may reduce probability calibration and ranking quality.

Based on the results in Table 5.5, Disagreement Weighted Cross-Entropy was selected as the default LNVA loss function. Although the non disagreement aware version achieves slightly higher accuracy in some configurations, this loss provides a more stable trade-off between accuracy and AUROC across different values of λ, suggesting improved learning capabilities in ranking quality. Fixing the loss function allows a more controlled analysis

of the effect of the weighting parameter λ, isolating its impact on LNVA prediction quality. Figure 5.1 illustrates the per-attribute LNVA f_1-score variation as a function of λ, highlighting how increasing task weighting influences both individual attribute performance and overall consistency across LNVA predictions.

FIGURE 5.1: Per-attribute f_1 score as a function of the task weighting parameter (λ) under fixed Disagreement Weighted Cross-Entropy supervision. Internal Structure and Calcification were excluded due to their extreme class imbalance see Appendix .3(Table A.4).

The results reveal a clear and structured effect of λ, increasing it systematically reshapes the learned representation in a controlled manner. At $\lambda = 0.1$, LNVA supervision has a limited influence on the model, resulting on an average f_1 score well below 0.3 and as λ increases, a more balanced regime emerges, where all attributes exhibit improved predictive performance. This enables the conclusion that increasing the importance we give

to the visual attributes, improves both the malignancy score (Table 5.5) and the LNVA's recognition (Figure 5.1). This begs the question, how far can we take lambda, being aware that in the semi-supervised step we will be summing another loss, without hurting malignancy prediction?

TABLE 5.6: Effects of increasing the LNVA loss weight (λ) on average malignancy classification performance and LNVA prediction quality (single-seed analysis). Fixed seed 42.

λ	**Accuracy**	**AUROC**	**LNVA** f_1
0.5	0.8649	0.8306	0.2951
1.0	0.8626	0.8020	0.3384
1.5	0.8590	0.7939	0.3232
2.0	0.8610	0.8074	**0.3711**
2.5	0.8695	0.7885	0.3372
3.0	0.8584	0.8246	0.3479
3.5	0.8445	0.8184	0.3314
4.0	0.8383	0.8157	0.3483

In Table 5.6, we see that attributes performance generally increases with λ, reaching its highest value at $\lambda = 2.0$. Malignancy performance also seems to degrade after the same value, so, we will be moving forward with Disagreement Weighted Cross-Entropy loss and a λ of 2.

Backbone Robustness

Before moving to semi-supervised learning, we tried to understand if the current model was the best option to learn both malignancy and visual attributes. Table 5.7 compares the performance of the three evaluated backbone architectures under identical conditions.

TABLE 5.7: Comparison of backbone architectures under the multi-task learning setting for malignancy

Backbone	**Accuracy**	**AUROC**
ResNet-18	$\mathbf{0.8581 \pm 0.0110}$	0.8054 ± 0.0129
DenseNet-18	0.8518 ± 0.0124	$\mathbf{0.8205 \pm 0.0098}$
Res2Net-18	0.8556 ± 0.0260	0.8091 ± 0.0257

While certain architectures achieved slightly stronger performance on specific metrics, like DenseNet on AUROC, the overall differences between backbones remained relatively small as the multitask framework itself is already robust. Consequently, further analysis focus on the visual attribute prediction tasks, so, to better understand how each backbone captured radiological semantic attributes, the following analysis examines the per-attribute LNVA f_1 scores through a heatmap-based comparison (Figure 5.2).

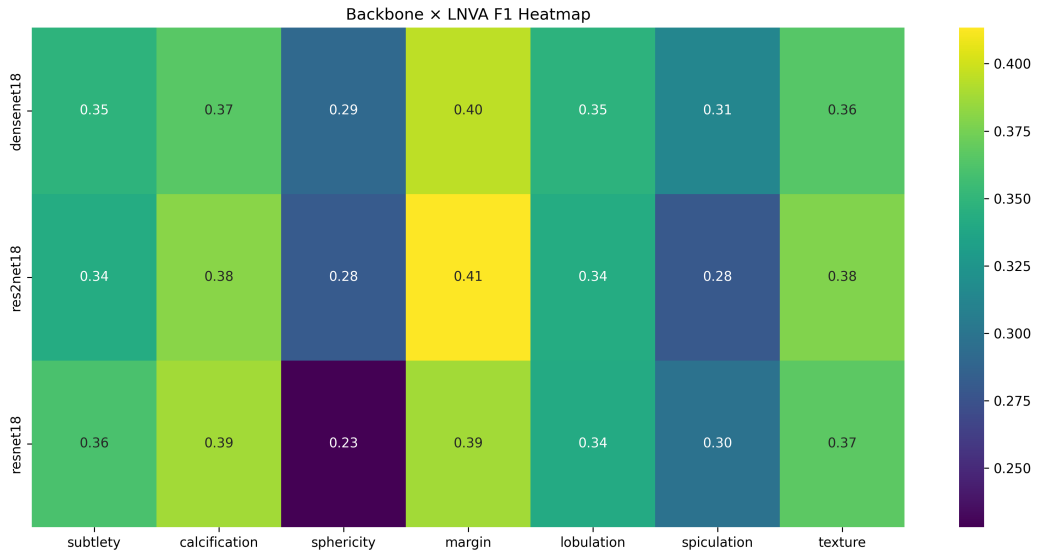


FIGURE 5.2: Per-attribute f_1 score heatmap comparing three backbones. Internal Structure was excluded due to its extreme class imbalance see Appendix .3(Table A.4).

Excluding Internal Structure due to its extreme class imbalance, DenseNet-18 achieved the highest f_1 scores for Lobulation, Sphericity and Spiculation, Res2Net-18 performed best on Margin and Texture, and ResNet-18 on Calcification and Subtlety. Although no architecture consistently dominated all attributes, DenseNet-18 provided the most balanced multitask performance and simultaneously achieved the highest malignancy classification AUROC. Given that LNVA prediction acts as an auxiliary task whose primary purpose is to improve malignancy classification and that attributes such as spiculation have been reported in the literature as strong indicators of malignancy [19], DenseNet-18 was selected for the subsequent semi-supervised experiments.

5.3 Semi-Supervised Learning

Building on the multi-task learning framework, this section investigates whether leveraging unlabeled data can further improve malignancy and lung nodule visual attribute

classification performance. All experiments follow the same evaluation protocol as previous sections, reporting mean ± standard deviation over five fixed random seeds. The semi-supervised setting builds upon the best multi-task configuration selected previously, resumed in Table 5.8, with the components held fixed unless otherwise specified.

TABLE 5.8: Base configuration for semi-supervised learning experiments.

Component	**Setting**
Backbone	DenseNet-18
Learning Rate	3×10^{-4}
Normalization	Batch Normalization
Malignancy Loss Function	Cross-Entropy
Malignancy Loss Weights	LUNA 0.5 — LIDC 0.5
LNVA Loss Function	Disagreement Weighted Cross-Entropy
LNVA Loss Weights	2

5.3.1 FixMatch-based consistency regularization

The semi-supervised learning strategy is based on FixMatch, which enforces prediction consistency between weakly and strongly augmented versions of unlabeled samples using high-confidence pseudo-labels (3.3). To analyze the sensitivity of the semi-supervised framework, a controlled hyperparameter sweep is performed over the FixMatch confidence threshold τ and the SSL loss weight α, as in Table 5.9.

TABLE 5.9: Semi-supervised learning experimental sweep configuration.

Component	**Values**
Pseudo-labeling threshold (τ)	$\{0.85,\ 0.90,\ 0.95\}$
SSL loss weight (α)	$\{0.5,\ 1,\ 2\}$
Random seeds	$\{42,\ 124,\ 168,\ 336,\ 672\}$

First, malignancy classification performance was analyzed to verify if the addition of the semi-supervised component does not negatively affect the primary task and only then, the LNVA prediction performance was examined to determine whether consistency regularization provides additional benefits for visual attribute learning.

TABLE 5.10: Average malignancy classification performance across LIDC-IDRI and LUNA25 datasets for different semi-supervised learning configurations.

Configuration		**Accuracy**	**AUROC**	**Δ AUROC (%)**
Threshold (th)	**Alpha (α)**			
0.85	0.5	0.8429 ± 0.0091	0.8260 ± 0.0106	$+0.55\%$
	1.0	0.8395 ± 0.0114	0.8262 ± 0.0096	$+0.57\%$
	2.0	0.8416 ± 0.0051	0.8308 ± 0.0062	$+1.03\%$
0.90	0.5	0.8459 ± 0.0095	0.8244 ± 0.0058	$+0.39\%$
	1.0	0.8436 ± 0.0047	0.8275 ± 0.0103	$+0.70\%$
	2.0	0.8403 ± 0.0122	0.8291 ± 0.0073	$+0.86\%$
0.95	0.5	0.8437 ± 0.0124	0.8185 ± 0.0078	-0.20%
	1.0	0.8496 ± 0.0059	0.8265 ± 0.0074	$+0.60\%$
	2.0	0.8463 ± 0.0042	0.8215 ± 0.0083	$+0.10\%$
(Multi-task Baseline)		$\mathbf{0.8518 \pm 0.0095}$	$\mathbf{0.8205 \pm 0.0087}$	-

Table 5.10 shows that semi-supervised training does not degrade malignancy classification performance relative to the multi-task supervised baseline. Across all nine configurations, AUROC consistently matches or slightly exceeds the baseline, with eight of nine settings yielding positive Δ AUROC (up to $+1.03\%$ at $\tau = 0.85$, $\alpha = 2.0$), while accuracy exhibits marginal degradation. Although the consistent AUROC gains may suggest a modestly better-calibrated decision boundary, the absence of a corresponding improvement in accuracy makes this interpretation inconclusive. Moreover, these marginal gains come at a substantial computational cost: average training time increases from approximately 32 minutes per run in the supervised multi-task setting to roughly 1h07m under FixMatch, more than doubling the training budget. Considered in isolation, the improvements in malignancy classification do not appear to justify this overhead.

It is also worth noting that malignancy is, in clinical practice, determined by a broader set of factors than the visual attributes alone, including patient-level information such as age, smoking history, occupational exposure and family history, none of which are available to the model. If the LNVAs capture only a fraction of the signal relevant to malignancy, the supervised multi-task objective may already be extracting most of the attribute-derived

information useful for this task. Under this view, further refining the LNVA predictions through semi-supervised learning could yield diminishing returns for malignancy classification, as the remaining performance gap is dominated by factors that are inherently inaccessible to an image-only model. This provides an additional interpretation for why the gains observed in Table 5.10 are limited, independently of the quality of the pseudo-labels themselves. The remainder of this section therefore examines whether the LNVA predictions benefit more substantially from the semi-supervised objective, which would provide a stronger rationale for adopting this framework.

FIGURE 5.3: Per-attribute f_1 score scores across semi-supervised configurations compared to the multi-task supervised baseline.

Figure 5.3 presents the per-attribute f_1 score scores for all semi-supervised configurations,

with the multi-task supervised baseline included as a reference column. The number of configurations that improve over the baseline for each attribute is shown in Table 5.11:

TABLE 5.11: Number of SSL configurations improving over the multi-task baseline, per attribute.

Attribute	**Baseline f_1**	**SSL wins**	**Best SSL f_1**
Calcification	0.365	6/9	0.413 ($\tau = 0.95$, $\alpha = 0.5$)
Margin	0.395	3/9	0.410 ($\tau = 0.95$, $\alpha = 1.0$)
Lobulation	0.347	4/9	0.387 ($\tau = 0.95$, $\alpha = 2.0$)
Texture	0.365	4/9	0.389 ($\tau = 0.95$, $\alpha = 1.0$)
Subtlety	0.348	1/9	0.360 ($\tau = 0.95$, $\alpha = 1.0$)
Spiculation	0.313	1/9	0.315 ($\tau = 0.90$, $\alpha = 0.5$)
Sphericity	0.291	0/9	0.282 ($\tau = 0.85$, $\alpha = 1.0$)
Average	**0.428**	**0/9**	0.427 ($\tau = 0.85$, $\alpha = 1.0$)

The results are mixed: semi-supervised learning improves LNVA prediction for some attributes, but not consistently across the board. Calcification exhibits the largest benefit, with six of the nine SSL configurations surpassing the baseline, with additional but less stable improvements observed for lobulation, texture and margin. In contrast, spiculation and subtlety show only marginal gains, while sphericity fails to improve under any evaluated configuration. At the aggregate level, however, no configuration exceeds the multi-task baseline average. In this context, the marginal degradation of accuracy shown in Table 5.10 may be partially explained by the nature of the LNVA signal itself. As the same attributes that show little to no gains from the SSL approach will be shown to be among the most predictive features for malignancy classification in Table 5.16, creating a setting where the pseudo-labels generated by the teacher model are both highly influential and potentially noisy.

Regarding the effect of the confidence threshold, a consistent trend emerges: $\bar{f}(0.95) \geq \bar{f}(0.90) \geq \bar{f}(0.85)$ across most attributes. This behavior is aligned with the increasing selectivity of pseudo-label filtering, where higher thresholds remove low-confidence predictions and retain only more reliable pseudo-labels. However, this comes at a significant cost in terms of unlabeled data utilization:

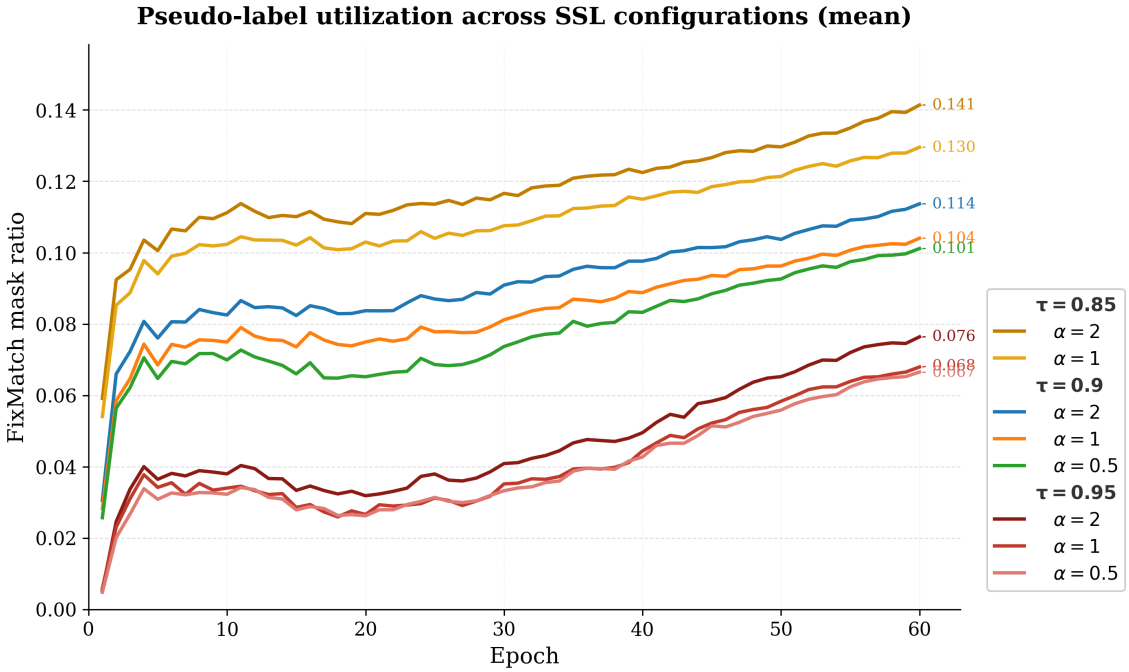


FIGURE 5.4: FixMatch mask ratio mean across LNVA attributes.

As shown in Figure 5.4, the effective mask ratio varies between 0.067 for $\tau = 0.95$ and 0.141 for $\tau = 0.85$, indicating that in the most conservative setting, more than 93.3% of unlabeled samples do not contribute to the SSL loss. This places the training in a low-utilization regime, where the contribution of the unsupervised objective is substantially reduced and the method behaves closer to a weak regularization of the supervised multi-task model rather than a fully effective semi-supervised learning framework. (see individual lnva mask ratio distribution Appendix .4 Fig. A.2)

This limited gain from unlabeled data must also be interpreted in the context of the underlying supervised model. As shown in Figure 5.1 from the Multi-Task Learning section, the baseline itself is not fully stable, with LNVA f_1 scores varying significantly across λ, despite relatively small changes in malignancy performance. This suggests that visual attribute learning is already a sensitive component of the joint optimization process, where performance is strongly dependent on the balance between malignancy and attribute losses.

Taken together, these factors suggest that the limited effectiveness of the semi-supervised framework is not solely due to the absence of useful unlabeled information, but rather to

a combination of (i) extremely low utilization of unlabeled data induced by high confidence thresholds and (ii) inherent ambiguity and instability in LNVA prediction.

Overall, these results indicate that FixMatch, in this setting, operates closer to a constrained regularization mechanism than a fully effective semi-supervised learning strategy, with its impact limited by low pseudo-label coverage and label ambiguity.

5.3.2 Threshold and weight scheduling

The analysis in Section 5.3.1 identified two structural limitations of the static FixMatch configuration: (i) the low pseudo-label utilization induced by high confidence thresholds, and (ii) the diminishing returns of further LNVA refinement once the supervised multi-task objective has already extracted most of the attribute-derived signal useful for malignancy. Table 5.10 also shows that the strongest Δ AUROC ($+1.03\%$) occurs at $\tau = 0.85$, $\alpha = 2.0$, precisely the regime that maximizes pseudo-label coverage. This motivates a more aggressive use of unlabeled data, provided it can be introduced safely. Before transitioning to UniMatch, four modifications to FixMatch are introduced jointly:

1. **Threshold scheduling.** τ is linearly annealed from 0.60 to 0.85 during training, admitting more pseudo-labels early on (when the model is less calibrated) and tightening as predictions become more reliable. This directly targets the low mask ratio identified in Figure 5.4.

2. **SSL weight ramp-up.** α is held at zero for 5 warm-up epochs and then linearly ramped over the next 30, preventing the unsupervised objective from interfering with representation learning before a meaningful supervised signal is acquired.

3. **Reduced LNVA weight.** α_{LNVA} is lowered from 2.0 to $\{1.0, 1.5\}$, freeing representational capacity from a near-saturated supervised signal so the FixMatch consistency term can contribute.

4. **Selective attribute supervision.** LNVA supervision is restricted to attributes with importance score above 1.0 in Table 5.16 (spiculation, sphericity, calcification, lobulation, subtlety), excluding margin, texture and internal structure, whose pseudo-labels are most likely to be noisy or weakly informative.

A sweep over $\alpha_{\text{FixMatch}} \in \{0.5, 1.0, 1.5\}$ and $\alpha_{\text{LNVA}} \in \{1.0, 1.5\}$ was conducted, with the remaining components fixed as in Table 5.8. Malignancy classification results are reported in Table 5.12.

TABLE 5.12: Average malignancy classification performance for the threshold-scheduled FixMatch configuration ($\tau_{\text{start}} = 0.60$, $\tau_{\text{end}} = 0.85$, warm-up $= 5$ epochs, ramp-up $= 30$ epochs), including Multi-task and SSL baselines.

Configuration		**Accuracy**	**AUROC**	**Δ AUROC (%)**	
FixMatch α	**LNVA α**			**vs MT**	**vs SSL**
0.5	1.0	0.8539 ± 0.0073	0.8146 ± 0.0064	-0.72%	-0.84%
	1.5	0.8381 ± 0.0069	0.8215 ± 0.0084	$+0.12\%$	0.00%
1.0	1.0	$\mathbf{0.8612 \pm 0.0090}$	0.8185 ± 0.0105	-0.24%	-0.37%
	1.5	0.8383 ± 0.0065	0.8262 ± 0.0109	$+0.69\%$	$+0.57\%$
1.5	1.0	0.8529 ± 0.0154	0.8238 ± 0.0104	$+0.40\%$	$+0.28\%$
	1.5	0.8488 ± 0.0155	$\mathbf{0.8273 \pm 0.0134}$	$+0.83\%$	$+0.71\%$
Multi-task Baseline		0.8518 ± 0.0095	0.8205 ± 0.0087	–	–
SSL Baseline		0.8463 ± 0.0042	0.8215 ± 0.0083	–	–

The results show no meaningful improvement over either the multi-task or the static FixMatch baselines. Most configurations remain within the variance range of the baselines, and no consistent pattern emerges across the α_{FixMatch} and α_{LNVA} combinations. Notably, the SSL weight ramp up does not appear to have had any beneficial effect on malignancy prediction, suggesting that the static FixMatch was not harming malignancy performance because of instability during early training, but rather because the signal it introduces is itself limited in this setting.

Examining the LNVA task more directly, Table 5.13 compares f_1 score scores for each attribute between the strongest static FixMatch configuration ($\tau = 0.95$, $\alpha_{\text{LNVA}} = 2.0$), the best scheduled variant ($\alpha_{\text{FixMatch}} = 1.5$, $\alpha_{\text{LNVA}} = 1.5$) and the Multi-task Baseline defined in 5.2, restricted to the five supervised attributes.

TABLE 5.13: Per-attribute f_1 score performance comparing the static FixMatch configuration ($\tau = 0.95$, $\alpha_{\text{LNVA}} = 2.0$) against the threshold-scheduled variant ($\tau_{\text{start}} = 0.60 \rightarrow \tau_{\text{end}} = 0.85$, $\alpha_{\text{LNVA}} = 1.5$), with Multi-task baseline included.

LNVA Attribute	**MT f_1**	**Static f_1**	**Scheduled f_1**	**Δ vs MT (%)**	**Δ vs Static (%)**
Subtlety	0.348	0.3389	0.3151	-9.48%	-7.02%
Calcification	0.365	0.3846	**0.4490**	$+23.01\%$	$+16.75\%$
Sphericity	0.291	0.2631	0.2090	-28.18%	-20.57%
Lobulation	0.347	0.3845	0.3221	-7.17%	-16.23%
Spiculation	0.313	0.3058	0.3013	-3.74%	-1.47%
Macro Average	0.3328	0.3354	0.3193	$-\mathbf{4.06}\%$	$-\mathbf{4.80}\%$

The scheduled variant delivers a notable gain for calcification ($+6.45\%$ over the static configuration, $+5.37\%$ over the multi-task baseline), but this improvement is not representative of the overall trend. Subtlety and spiculation remain essentially unchanged, while sphericity degrades substantially under both SSL configurations relative to the multi-task baseline, and lobulation, despite improving over the baseline with the static configuration, regresses under scheduling. At the macro level, the scheduled variant underperforms both the multi-task baseline and the static FixMatch configuration, with a macro average f_1 of 0.3193 against 0.3354 and 0.3328 respectively.

The root cause of this behaviour is visible in Figure 5.5. Despite the lower starting threshold, the mean mask ratio never exceeds 0.12 throughout training, and fails to show the sustained growth that would be expected if the model were progressively generating more reliable pseudo-labels. The interplay between the ramp-up of the SSL weight and the tightening threshold produces a roughly stable effective contribution, but the underlying problem remains: for multi-class LNVA attributes with five ordinal categories, the predicted probability mass is distributed across classes in a way that rarely yields a confident winner above any reasonable threshold. Reducing the number of LNVAs under supervision to five does not resolve this, as the attributes with the highest importance scores (spiculation, sphericity) are precisely those where the model's f_1 is lowest, making their pseudo-labels the least reliable.

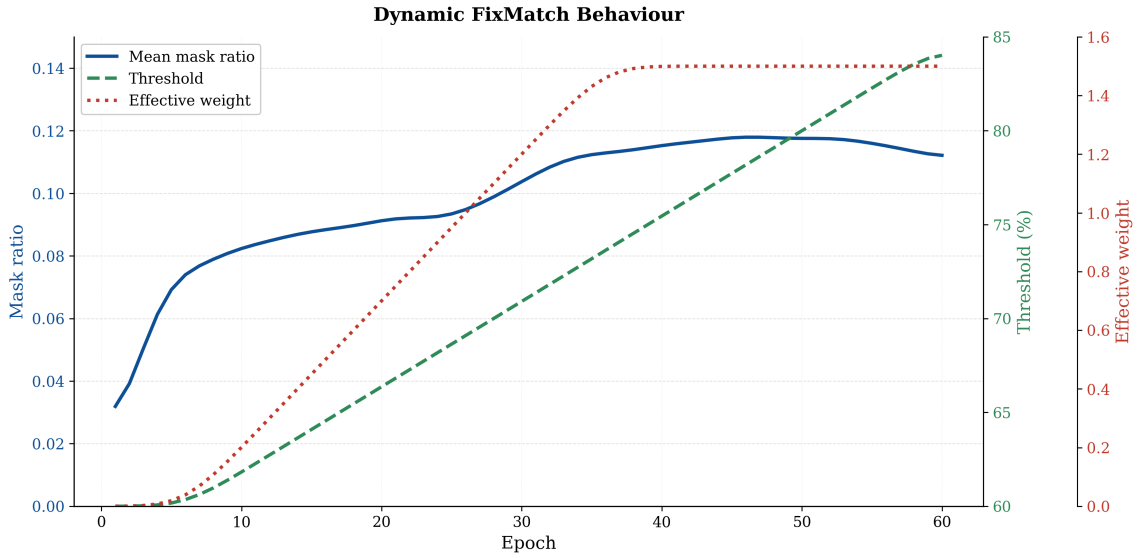


FIGURE 5.5: Mean mask ratio, confidence threshold and effective SSL weight under the scheduled FixMatch configuration ($\tau_{\text{start}} = 0.60 \rightarrow \tau_{\text{end}} = 0.85$, $\alpha_{\text{LNVA}} = 1.5$)

Taken together, these results suggest that the limitations of FixMatch in this setting are structural rather than a consequence of suboptimal hyperparameters: the combination of five-class ordinal prediction targets, domain shift between LIDC-IDRI and LUNA25, and the inherent difficulty of the LNVA task produces a pseudo-label quality ceiling that neither static nor scheduled thresholding can overcome. This motivates the transition to UniMatch, which addresses some of these limitations through a dual-stream consistency strategy that does not rely exclusively on hard thresholding for pseudo-label selection.

5.3.3 UniMatch (dual-stream consistency)

The analysis in Section 5.3.2 established that the structural ceiling on FixMatch performance in this setting is not primarily a function of hyperparameter choice, but of the difficulty of generating reliable pseudo-labels for five-class ordinal LNVA targets under domain shift. Hard confidence thresholding is ill-suited to this regime: when predicted probability mass is distributed across multiple ordinal categories, any single class rarely accumulates sufficient confidence to cross a fixed threshold, regardless of how that threshold is tuned. UniMatch addresses this limitation by replacing the single strongly-augmented consistency target with a dual-stream formulation. For each unlabeled sample, two independently sampled strong augmentations and a feature-dropout view are generated, and the consistency loss is computed as the mean cross-entropy between the

weak-view pseudo-label and all three targets (2). This design has two practical consequences. First, by averaging the loss over three views rather than one, the gradient estimate from unlabeled samples has lower variance, which can stabilize training when pseudo-label quality is heterogeneous. Second, the feature-dropout view introduces perturbation directly in the latent space, decoupling the consistency signal from purely pixel-level variation and making it more robust to the input distribution shift between datasets.

UniMatch was evaluated under the same base configuration as the scheduled FixMatch variant ($\alpha_{\text{FixMatch}} = 1.5$, $\alpha_{\text{LNVA}} = 1.5$), varying only the feature dropout rate applied to the encoder representation. Table 5.14 reports malignancy classification performance across dropout values.

TABLE 5.14: Average malignancy classification performance for different UniMatch dropout values using $\alpha_{\text{FixMatch}} = 1.5$ and $\alpha_{\text{LNVA}} = 1.5$. Accuracy and AUROC correspond to the mean performance across LIDC-IDRI and LUNA25. The AUROC improvement is reported relative to the threshold-scheduled FixMatch baseline.

UniMatch Dropout	Accuracy	AUROC	Δ FixMatch Schedule (%)
0.0	$\mathbf{0.8621 \pm 0.0196}$	0.8115 ± 0.0205	-1.91%
0.1	0.8381 ± 0.0169	0.8169 ± 0.0147	-1.26%
0.3	0.8373 ± 0.0122	$\mathbf{0.8256 \pm 0.0064}$	-0.21%
0.5	0.8585 ± 0.0164	0.8152 ± 0.0179	-1.46%
FixMatch Schedule Baseline	0.8488 ± 0.0155	0.8273 ± 0.0134	–

UniMatch does not improve malignancy classification relative to the scheduled FixMatch baseline. All dropout variants fall below the FixMatch Schedule AUROC of 0.8273, with the closest configuration (dropout $= 0.3$) still registering a -0.21% gap. This pattern is consistent with the general finding that gains in malignancy performance from the semi-supervised branch are marginal and within the range of variance across seeds, suggesting that the dual-stream extension does not resolve the underlying bottleneck on the primary task. The LNVA results present a more nuanced picture. Table 5.15 compares per-attribute f_1 scores for dropout values of 0.1 and 0.5 against the scheduled FixMatch baseline.

TABLE 5.15: Per-LNVA f_1 score performance comparing UniMatch dropout variants (0.1 and 0.5) against the threshold-scheduled FixMatch baseline ($\tau_{\text{start}} = 0.60 \rightarrow \tau_{\text{end}} = 0.85$, $\alpha_{\text{LNVA}} = 1.5$). The Δ columns report relative improvement over the FixMatch Schedule baseline.

LNVA Attribute	**Dropout 0.1**	**Dropout 0.5**	**FixMatch Schedule**	**Δ FixMatch Schedule (%)**	
				0.1	**0.5**
Subtlety	0.307	0.309	0.3151	-2.57%	-1.94%
Calcification	0.439	0.462	0.4490	-2.23%	$+2.90\%$
Sphericity	0.200	0.210	0.2090	-4.31%	$+0.48\%$
Lobulation	0.339	0.332	0.3221	$+5.25\%$	$+3.08\%$
Spiculation	0.314	0.312	0.3013	$+4.22\%$	$+3.56\%$
Macro Average	0.3198	0.3250	0.3193	$+0.16\%$	$+1.77\%$

At the macro level, UniMatch with dropout $= 0.5$ marginally outperforms both the scheduled FixMatch baseline ($+1.77\%$) and the dropout $= 0.1$ variant ($+0.16\%$). Lobulation and spiculation show consistent improvements across both dropout values, and calcification benefits further under higher dropout, while subtlety and sphericity remain below baseline in both configurations. These attribute-level results echo the pattern already observed in the static FixMatch analysis: the attributes that benefit most from consistency regularization are those with moderate class imbalance and sufficient pseudo-label coverage, while the most imbalanced and shape-sensitive attributes remain resistant to improvement across all semi-supervised formulations. However, the mask ratio for UniMatch does not meaningfully improve over its FixMatch counterpart. Since UniMatch retains the same confidence-thresholded pseudo-label mechanism for mask generation, the fundamental utilization bottleneck identified in Section 5.3 persists: the five-class ordinal structure of the LNVA targets prevents any single class from accumulating sufficient predicted probability to clear the threshold consistently, and the dual-stream design does not address this directly (Figure 5.6). What UniMatch provides is a more diverse and lower-variance consistency signal for those samples that do pass the threshold, but it cannot increase the number of samples that do so. The marginal macro-level gains at dropout $= 0.5$ are therefore best interpreted as a modest improvement in the quality of the consistency gradient rather than a meaningful expansion of pseudo-label coverage.

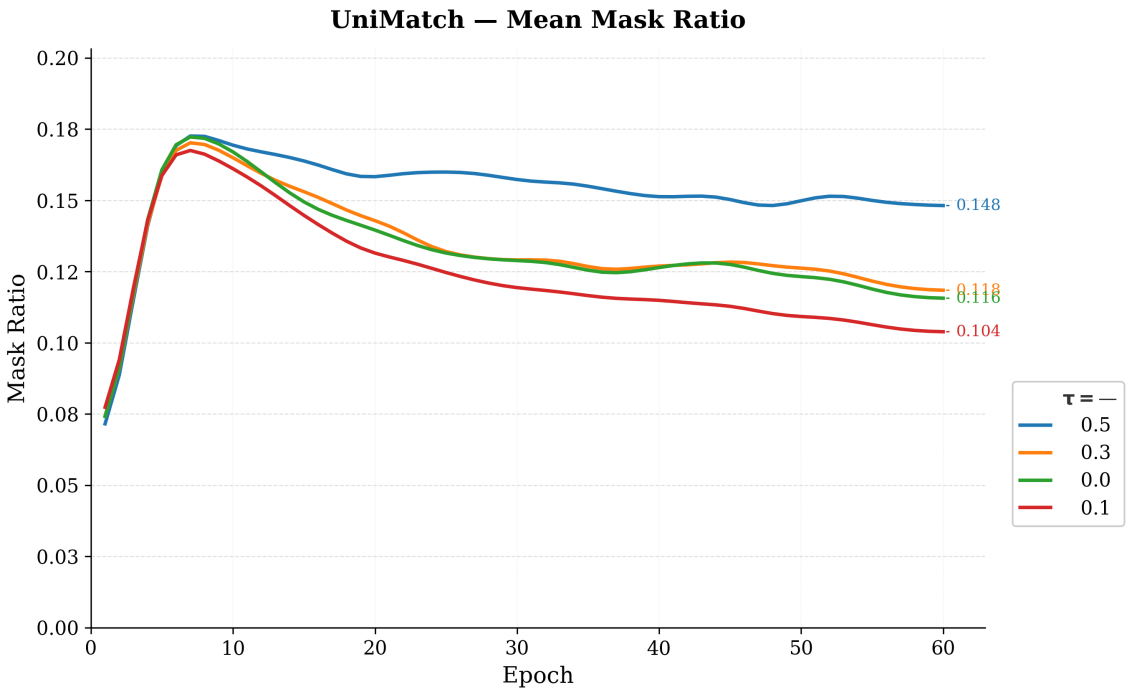


FIGURE 5.6: Unimatch mean mask ration

Taken together, the UniMatch results confirm that the limitations of the semi-supervised framework in this setting are structural. Neither the augmentation diversity introduced by dual-stream consistency, nor the latent-space perturbation from feature dropout, nor the scheduling strategies explored in Section 5.3.2 are sufficient to overcome the combination of low pseudo-label utilization and inherent LNVA label ambiguity. The bottleneck lies upstream: in the difficulty of generating reliable high-confidence pseudo-labels for multi-class ordinal targets under domain shift, a challenge that cannot be resolved through consistency design alone without addressing the label space or the domain gap directly.

5.4 Interpretability and Attribute Contribution

The interpretability analysis is conducted on the semi-supervised model obtained in the phase 5.3.1, which represents the canonical FixMatch-based configuration of the proposed framework. The scheduled variant introduced in Phase 5.3.2 threshold and weight scheduling and the UniMatch (dual-stream consistency) in phase 5.3.3 are treated as an exploratory extension and are not included in this analysis. This section investigates the internal behavior of the proposed framework through spatial explanations and then a post-hoc analysis of LNVA contributions to malignancy prediction. The goal is to assess

how different training strategies affect both the model's spatial focus and the relative importance of learned visual attributes.

Spatial explanations

To investigate whether the model learns malignancy primarily from the nodule itself or from surrounding contextual structures, and how this behavior is affected by the introduction of multi-task learning, Grad-CAM visualizations were analyzed across four network depths, comparing activation maps between the purely supervised model and the LNVA-augmented semi-supervised model. While Grad-CAM explanations at deeper layers are inherently less interpretable from a human perspective, they remain useful for assessing relative shifts in spatial attention across configurations. Figure 5.7 illustrates the evolution of activation patterns across four layers for representative pulmonary nodule samples. Each row corresponds to a specific case, annotated with the prediction outcome of both models in the format *[Supervised | Semi-Supervised]*.

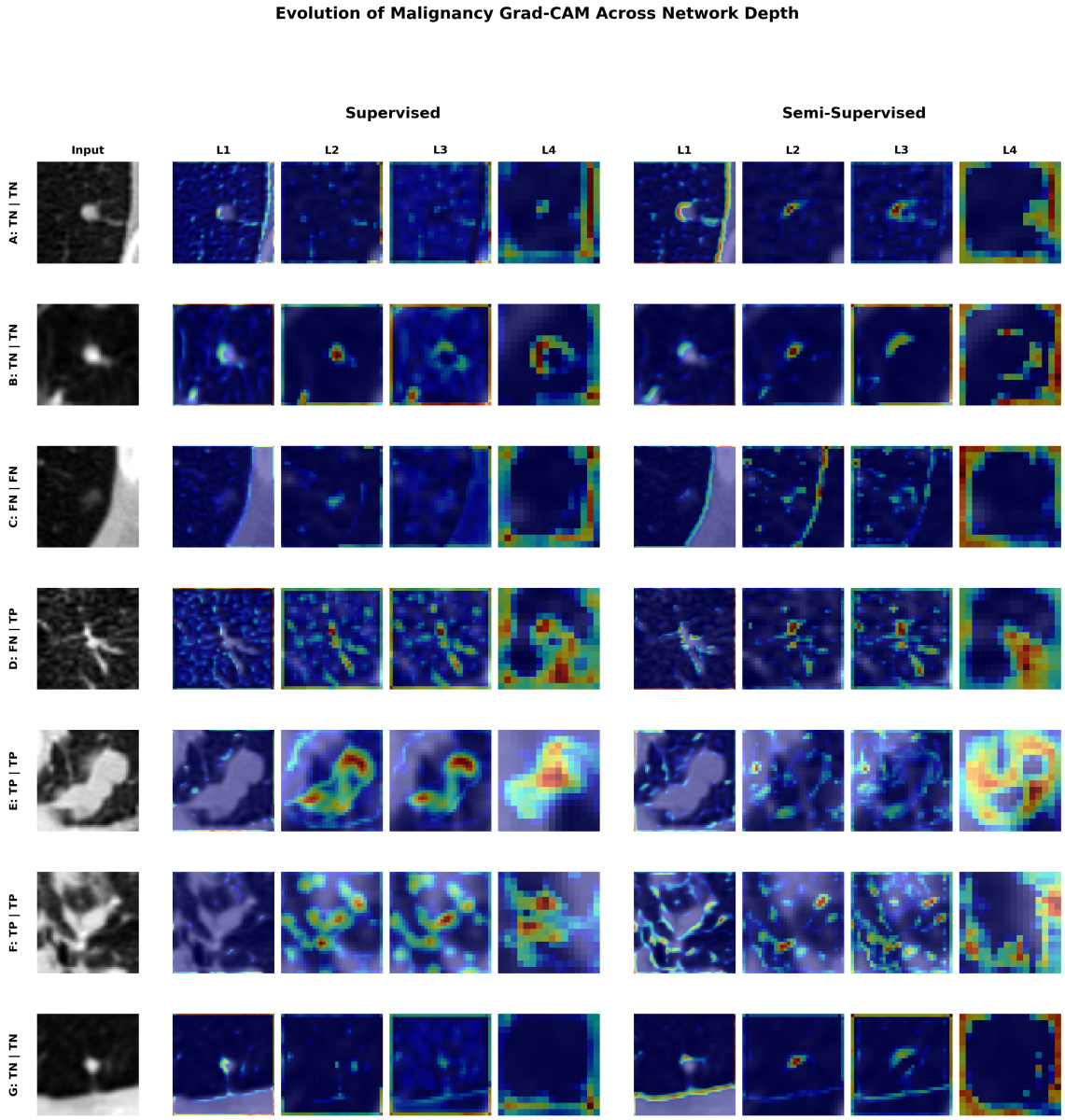


FIGURE 5.7: Evolution of malignancy Grad-CAM activations across network depth. Comparison between supervised (left) and semi-supervised (right) models across four layers. Each row corresponds to a specific pulmonary nodule sample, identified on the left by its classification outcome formatted as [Supervised Outcome | Semi-Supervised Outcome].

Across samples, a consistent qualitative difference emerges between the two models. The supervised baseline tends to focus more narrowly on the nodule region when confident in its prediction (rows A, B, E and F), whereas in uncertain cases its attention becomes more diffuse and partially shifts to surrounding regions (rows E and F). In contrast, the semi-supervised model exhibits a more structured and spatially distributed activation pattern, frequently extending its focus beyond the nodule boundary, possibly indicating that the

semi-supervised model is incorporating contextual cues from the surrounding lung characteristics, potentially as a complementary signal for malignancy assessment. This would make sense for radiological attributes such as spiculation or lobulation, which inherently dependency on boundary morphology and local context.

Overall, these observations suggest that the inclusion of visual attributes learning does not merely refine classification performance, but also induces a shift in the learned representation from purely object-centric reasoning toward a more context-aware decision process. While both models achieve comparable quantitative performance, their internal decision strategies appear to differ in a interpretable manner.

Now we will look exclusively at the semi-supervised model and examine how LNVA supervision influences spatial attention patterns in the final convolutional representation. Figure 5.8 presents a transposed organization of Grad-CAM maps, structured by classification outcome (TP, TN, FP, FN) and by LNVA attribute head.

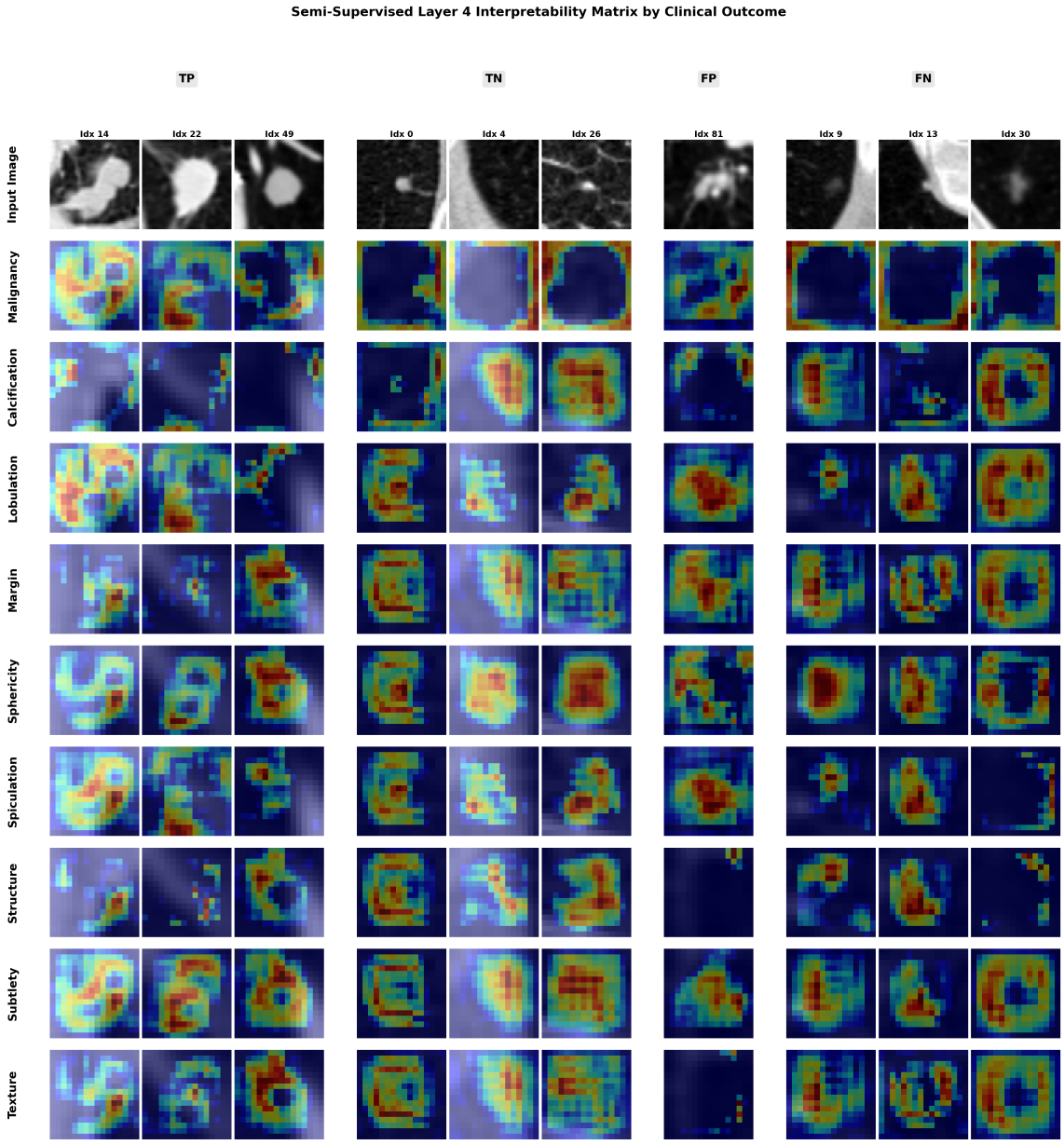


FIGURE 5.8: Transposed matrix of Grad-CAM activation maps (Layer 4) from the semi-supervised model, structured by classification outcome categories (TP, TN, FP, FN). Each column corresponds to a distinct lung nodule sample, headed by its input image. Rows correspond to different attribute heads.

A first consistent observation is that visual attribute supervision induces stable and non-random activation across all samples, regardless of the final malignancy prediction. In other words, attribute heads remain visually responsive across cases, suggesting that attribute learning is not suppressed by the primary malignancy objective, but rather coexists with it in the shared representation space. A second and more relevant pattern emerges when conditioning on malignancy outcome. In true positive and false positive cases, the model tends to concentrate activations more strongly within the nodule core region,

but in true negative and false negative cases, activation maps are generally more diffuse and extend towards surrounding lung tissue and boundary regions. This suggests that the model associates more localized internal representations with malignant predictions, while benign predictions are more strongly influenced by contextual or peripheral information. Although this behavior is not strictly causal, it is consistent with the hypothesis that the model encodes malignancy as a combination of intra-nodular appearance and boundary irregularity cues, rather than relying solely on a single spatial region.

When analyzing LNVA heads specifically, a further distinction appears. Attributes such as calcification, margin and internal structure tend to produce slightly more spatially dispersed activation patterns. This is consistent with their clinical definition, which frequently depends on contextual or boundary-level characteristics rather than internal texture alone. In contrast, attributes such as spiculation, lobulation and subtlety exhibit a more concentrated activation around the nodule edge and internal region, reflecting their dependence on fine-grained shape deformation and local morphological variation.

It is also important to note that the presented qualitative analysis contains a limited number of false positive cases, with only a single representative FP example available in this visualization. While this prevents any systematic conclusions about false positive behavior, the observed pattern remains informative when interpreted jointly with true positive cases, *i.e.*, as the set of samples predicted as malignant by the model. Taking this in consideration, these results suggest that multi-task learning encourages the model to distribute attention across both intra-nodular and peri-nodular regions in a structured manner, rather than focusing exclusively on the lesion core. This supports the hypothesis that multi-task learning does not only improve predictive performance, but also shapes a more anatomically consistent representation of pulmonary nodules.

Quantitative relationship between attributes and malignancy

To understand which visual attributes are most predictive of malignancy, we computed a regression-based importance score derived from their prediction heads (Chapter 4, Section 4.9). Table 5.16 summarizes the ranking results.

TABLE 5.16: LNVA attribute importance ranking based on logistic regression coefficient norm with respect to malignancy prediction.

LNVA Attribute	Importance Score
Spiculation	2.4182
Sphericity	1.7236
Calcification	1.2674
Lobulation	1.2009
Subtlety	1.0845
Texture	0.4076
Margin	0.2061
Structure	0.0000

The structure attribute consistently yields a near-zero importance score, which is also consistent with its extreme class imbalance in the dataset, see Appendix .3 (Table A.4).

The other results reveal a separation between shape-related attributes and appearance-based descriptors. Attributes such as spiculation, sphericity and lobulation dominate the ranking, while attributes such as texture, margin and internal structure exhibit substantially lower importance scores, indicating that geometric deformation and boundary irregularity carry a stronger predictive signal for malignancy classification, with the only exception being calcification, which is somewhat surprising given its relatively high importance despite being a highly imbalanced attribute and typically less consistently present across samples. This suggests that calcification may still act as a discriminative contextual signal in distinguishing between benign and malignant cases.

5.5 Discussion

The experimental protocol followed in this work was designed as a deliberate progression rather than a parallel comparison. Moving through supervised, multi-task and semi-supervised learning as sequential stages, each building on the best configuration selected in the previous one, served a purpose that goes beyond model selection: it imposed a structured understanding of the problem at each level before adding complexity. This

staged approach revealed that certain challenges, such as the sensitivity of LNVA learning to task weighting or the instability of consistency regularization under low pseudo-label coverage, would have been difficult to diagnose in a single consolidated experiment. In this sense, the experimental pipeline itself constitutes a methodological contribution, enabling a more grounded interpretation of the results that follow.

What worked

The most consistent finding across all phases is that Multi-Task Learning does not degrade malignancy classification performance relative to the Supervised Baseline. In fact, several configurations matched or slightly exceeded the baseline on AUROC, suggesting that auxiliary supervision acts as a beneficial inductive bias rather than a competing objective. This is a non-trivial result in multi-task learning, where auxiliary losses frequently introduce conflicting gradient directions that can harm the primary task, particularly when the auxiliary signal is noisy. The fact that this degradation did not materialize here supports the hypothesis that LNVA annotations, despite their inter-observer variability, carry structured information compatible with the shared feature space required for malignancy prediction.

Regarding the LNVAs themselves, the results demonstrate that the model learns non-trivial attribute representations even under the challenging conditions imposed by class imbalance and annotation noise. The clear dependence of per-attribute f_1 on the task weighting parameter λ confirms that the model is genuinely capable of learning these attributes when sufficient optimization pressure is applied. The Grad-CAM analysis further supports this, showing that individual attribute heads produce spatially distinct and anatomically interpretable activations, particularly for shape-sensitive attributes such as spiculation and lobulation, which concentrate their activation around nodule boundaries in a manner consistent with their clinical definitions.

Semi-supervised learning underperformed

The limited effectiveness of the semi-supervised extension can be explained by three interacting factors. The most quantitatively significant is the low pseudo-label utilization rate. As shown in Appendix .3 (Figure 5.4) the effective mask ratio for $\tau = 0.95$ averages around 0.06, meaning that fewer than one in fifteen unlabeled samples contribute to the

SSL loss at any given training step. This places the method in a regime where the unsupervised objective behaves more like a weak periodic regularizer than a meaningful second supervision source, and is consistent with the marginal and inconsistent improvements observed across LNVA attributes.

A second contributing factor is the domain shift between LIDC-IDRI and LUNA25. The attribute heads are trained exclusively on LIDC-IDRI, where annotations follow the specific scoring conventions of the four-radiologist protocol, while LUNA25 samples come from a different acquisition context, with different patient populations and imaging conditions. Pseudo-labels generated for these samples are therefore produced by an encoder that has never seen comparable data in a supervised setting, which limits the reliability of even high-confidence predictions.

Finally, there is the question of augmentation design. The strong augmentation pipeline used in FixMatch includes Random Erasing with a spatial range of 2–20% of the image area. While this constraint was designed to preserve global lesion morphology, it may still interfere with the fine-grained boundary and texture features that the LNVA heads rely on. This is particularly problematic for attributes whose classification already had limited performance in the multi-task setting, and could explain why attributes such as sphericity, which was already the weakest in multi-task learning, showed no improvement under any SSL configuration. A targeted ablation of the augmentation strategy would be needed to isolate this effect.

Model selection and evaluation

A recurring observation throughout this work is that accuracy and AUROC, while informative, provide an incomplete picture of model behavior, particularly in medical imaging settings where the cost of false negatives is substantially higher than the cost of false positives. A model that achieves strong AUROC by maintaining good class ranking may still produce unacceptable false negative rates in a clinical deployment scenario, where missing a malignant nodule has direct consequences for patient outcome. The evaluation framework used here does not fully reflect this asymmetry. Sensitivity and false negative rate should carry greater weight in model selection decisions, and future experiments should incorporate these metrics as primary criteria rather than secondary observations.

A related limitation concerns the absence of formal statistical testing. The comparisons reported across configurations rely on mean and standard deviation over five seeds, which provides some indication of stability but does not establish statistical significance. Pairwise tests such as the Wilcoxon signed-rank test, applied across seed-level results, would allow stronger claims about whether observed differences reflect genuine performance improvements or fall within the range of random variation. This is particularly relevant for the SSL results, where several configurations show improvements over the baseline for specific attributes that are numerically small and may not be statistically reliable.

The LNVA importance analysis reported in Table 5.16 offers a useful ordering of attribute predictiveness, but the logistic regression framework used to derive it has a known limitation: it evaluates each attribute independently, without accounting for redundancy or complementarity between attributes. A forward selection procedure, which sequentially adds attributes based on their marginal contribution to a jointly trained classifier would provide a more principled characterization of which attribute combinations are most informative for malignancy prediction. This approach has been shown to yield more interpretable and robust importance rankings in related multi-attribute prediction settings [19], and represents a natural extension of the post-hoc analysis conducted here.

Finally, the fixed weighting scheme used throughout this work, where λ values for each loss component are set prior to training and held constant, introduces a structural tension. The optimal balance between malignancy, LNVA and SSL losses is not known in advance and may shift as the model converges. Selecting $\lambda = 2$ for the LNVA loss based on supervised multi-task experiments does not guarantee that this value remains appropriate once the FixMatch objective is added, as the total loss landscape changes with the introduction of a new term. Uncertainty-aware multi-task weighting, in which each loss component is associated with a learnable noise parameter that dynamically adjusts its contribution during training, offers a principled solution to this problem and represents a direct methodological improvement over the fixed-weight formulation used here.

Chapter 6

Conclusion

This thesis investigated whether clinically meaningful visual attributes could serve as a structured supervision signal to improve lung nodule malignancy classification, and whether semi-supervised consistency regularization could extend that signal to unlabeled data. The experimental protocol was deliberately staged, progressing from a supervised baseline through multi-task learning and into semi-supervised learning, before closing with an interpretability analysis. This progression allowed each component to be evaluated in isolation, making the contribution of each design choice traceable rather than confounded by simultaneous changes.

Can nodule visual attribute supervision improve malignancy classification performance?

The answer is cautiously affirmative. Multi-task learning with LNVA supervision did not degrade malignancy classification relative to the supervised baseline, the configurations matched or slightly exceeded it on AUROC (5.5). This is a non-trivial result: in multi-task settings with noisy auxiliary signals, auxiliary losses frequently introduce conflicting gradients that harm the primary task. This degradation did not materialize, suggesting that visual attribute annotations, despite their inter-observer variability and class imbalance, carry structured information that is compatible with the shared feature space required for malignancy prediction.

The dependence of per-attribute f_1-score on the task weighting parameter λ confirmed that the model is genuinely responsive to attribute supervision when sufficient optimization pressure is applied (5.1), and that gains in attribute learning do not come at the cost

of malignancy performance up to a certain threshold. The selection of DenseNet-18 as the final backbone further reinforced this, as it provided the most balanced multi-task performance while simultaneously achieving the highest malignancy AUROC among the evaluated architectures (5.7).

It is worth noting, however, that the gains observed here are modest, and that this is perhaps expected. Malignancy is a clinical determination that extends well beyond what is visible in a single CT patch: patient age, smoking history, family history, and longitudinal nodule growth, none of which are available to the model, all contribute to real-world diagnostic decisions. Visual attributes capture only a fraction of this signal, and so a ceiling on how much attribute-guided representations can improve image-only malignancy classification is inherent to the problem formulation itself. Within that ceiling, auxiliary LNVA supervision acts as a beneficial inductive bias rather than a competing objective, lending support to the hypothesis that attribute-guided representations encode information that is complementary to, and not in conflict with, the primary classification signal.

Can semi-supervised learning improve LNVA prediction by leveraging samples without attribute annotations?

The answer is, for the most part, negative. At least under the conditions explored in this work. The FixMatch-based semi-supervised framework produced marginal and inconsistent improvements in attribute prediction across the evaluated configurations, with no configuration surpassing the multi-task baseline at the macro-average level (5.10, 5.11).

To understand why, it helps to trace the full chain of difficulty that the semi-supervised objective faces in this setting. The supervised multi-task model, which serves as the teacher generating pseudo-labels, itself achieves modest f_1 scores on the visual attributes (5.2). This is unsurprising given the nature of the task: the ground truth label for each attribute is the median across multiple radiologist annotations, introducing label noise from the outset. Each attribute is then a five-class, some ordinal, some categorical, target, meaning that the predicted probability mass is distributed across categories in a way that rarely yields a confident winner above any reasonable threshold. The result is an effective mask ratio below 0.15 across all configurations (5.4), placing the semi-supervised objective in a regime closer to weak regularization than genuine second supervision. And all of this

is framed as an auxiliary task, secondary to malignancy classification, which further constrains the representational capacity available for attribute learning. Taken together, these compounding factors make the limited effectiveness of the semi-supervised extension unsurprising in retrospect, even if it was not immediately obvious at the outset.

Threshold scheduling partially addressed the utilization problem for calcification, but failed to generalize across the remaining attributes, and the macro-average f_1 of the scheduled variant underperformed both the multi-task baseline and the static FixMatch configuration (**??**). UniMatch, by replacing hard thresholding with dual-stream consistency, provided a modest improvement at the macro level, but remained insufficient to consistently outperform the supervised multi-task baseline (5.15). These results suggest that the ceiling on semi-supervised attribute learning in this setting is not primarily a function of hyperparameter choice, but of the difficulty of generating reliable pseudo-labels for ordinal, multi-class targets under domain shift, a structural limitation that neither FixMatch nor UniMatch fully resolves within the framework evaluated here.

Does attribute-guided learning influence the spatial and semantic behavior of the model in an interpretable manner?

Yes, and this is perhaps the most consistent finding of this work. The Grad-CAM analysis revealed that LNVA supervision induces a qualitative shift in the spatial attention of the model, moving away from the narrowly nodule-focused activations of the supervised baseline toward a more distributed pattern that extends into the boundary and peri-nodular regions (5.7). This shift is anatomically consistent with attributes such as spiculation and lobulation, whose clinical definitions depend precisely on boundary morphology and local context, and suggests that the model is not merely learning to classify, but learning to attend to the image regions that a radiologist would consider relevant.

Beyond spatial attention, the individual attribute heads produced stable and non-random activations across all samples and outcome categories, confirming that attribute learning coexists with the primary malignancy objective rather than being suppressed by it. The quantitative importance analysis further revealed a clear separation between shape-related attributes, namely spiculation, sphericity, and lobulation, and appearance-based descriptors such as texture and margin, with the former carrying substantially stronger

predictive signal for malignancy (5.8). Calcification represented the notable exception, exhibiting higher importance than its class imbalance and clinical prevalence would suggest, indicating that it may still serve as a discriminative contextual signal in the binary malignancy setting. Taken together, these findings support the hypothesis that multi-task learning does not merely refine predictive performance, but actively shapes a more anatomically grounded and semantically structured internal representation of pulmonary nodules.

6.1 Limitations

Domain shift between datasets. LIDC-IDRI and LUNA25 differ in acquisition protocol, patient population, and annotation methodology. LNVA pseudo-labels for LUNA25 are generated by a model trained exclusively on LIDC-IDRI, which limits their reliability even at high confidence thresholds.

Malignancy label binarization. Malignancy scores in LIDC-IDRI are provided on an ordinal scale from 1 to 5, annotated by four radiologists. In this work, labels are binarized by assigning nodules with median score ≤ 2 as benign and those with median score ≥ 4 as malignant, while ambiguous cases with median score of 3 are excluded from malignancy evaluation. This threshold-based binarization discards a substantial portion of the data and may not capture the full continuum of malignancy risk, potentially introducing a selection bias toward more unambiguous cases.

Fixed loss weighting. The weighting parameters λ for each loss component were determined through staged grid search and held fixed throughout training. This approach does not account for the dynamic shifts in the loss landscape introduced when the SSL objective is added to the multi-task configuration. Uncertainty-aware multi-task weighting, in which each loss is associated with a learnable homoscedastic uncertainty parameter, offers a principled alternative and represents a direct improvement over the fixed-weight formulation used here.

Absence of formal statistical testing. Comparisons across configurations rely on mean and standard deviation over five random seeds. While this provides a measure of stability,

it does not establish statistical significance. Pairwise tests such as the Wilcoxon signed-rank test, applied across seed-level results, would allow stronger claims to be made, particularly for the SSL results where numerical improvements over the baseline are small.

Independent attribute importance analysis. The logistic regression framework used to assess LNVA predictiveness evaluates each attribute independently, without accounting for redundancy or complementarity between attributes. A forward selection procedure, sequentially adding attributes based on their marginal contribution to a jointly trained classifier, would yield a more principled characterization of which attribute combinations are most informative for malignancy prediction.

Evaluation metric asymmetry. AUROC and accuracy provide an incomplete picture in a clinical deployment context, where the cost of false negatives substantially exceeds the cost of false positives. Sensitivity and false negative rate should carry greater weight in future model selection criteria to better reflect the clinical consequences of missed malignancy detections.

6.2 Future Work

Several directions extend naturally from the findings of this thesis. First, the incorporation of uncertainty-aware multi-task weighting would address the fixed-loss-balance limitation identified above, allowing the optimization to dynamically adapt to the relative difficulty of each task throughout training. Second, adaptive confidence thresholds in the FixMatch branch - rather than the static and pre-defined τ values explored here - could improve pseudo-label utilization without introducing noise, particularly in the early training stages where the attribute heads are still converging. Third, reducing the cardinality of the LNVA label space represents a promising direction for improving pseudo-label quality in the semi-supervised setting. The current five-class ordinal formulation distributes probability mass across categories in a way that rarely yields confident predictions, directly limiting mask utilization. Collapsing each attribute into a binary or three-class representation would concentrate probability mass and make the FixMatch confidence threshold more effective, at the cost of some label granularity that may not be clinically necessary for the purposes of malignancy prediction. Fourth, a forward attribute selection analysis would complement the importance ranking reported here (5.16), providing

a more interpretable and statistically grounded characterization of which LNVA combinations maximally support malignancy prediction. Fifth, incorporating patient-level clinical context beyond image-derived features, such as age, smoking history, longitudinal nodule growth and family history, could substantially improve malignancy prediction, since these factors are routinely used in clinical decision-making and are entirely absent from the current image-only framework. Finally, validating the framework on external CT datasets beyond LIDC-IDRI and LUNA25 would assess the generalizability of the learned representations and the robustness of the multi-task design to variation in acquisition protocols and clinical settings.

Appendices

.1 LIDC-IDRI Feature Definitions

TABLE A.1: Names, descriptions, and score definitions of LIDC-IDRI semantic features (LNVAs), adapted from [22].

Name	Description	Score Definitions
Subtlety	Difficulty in detecting the nodule	1. Extremely Subtle 2. Moderately Subtle 3. Fairly Subtle 4. Moderately Obvious 5. Obvious
Internal structure	Internal composition of the nodule	1. Soft Tissue 2. Fluid 3. Fat 4. Air
Calcification	Calcification pattern of the nodule	1. Popcorn 2. Laminated 3. Solid 4. Non-central 5. Central 6. Absent
Sphericity	3D shape of the nodule in terms of its roundness	1. Linear 2. Ovoid/Linear 3. Ovoid 4. Ovoid/Round 5. Round

Continued on next page

Name	Description	Score Definitions
Margin	Nodule margin definition	1. Poorly Defined 2. Near Poorly Defined 3. Medium Margin 4. Near Sharp 5. Sharp
Lobulation	Degree of nodule lobulation	1. No Lobulation 2. Nearly No Lobulation 3. Medium Lobulation 4. Near Marked Lobulation 5. Marked Lobulation
Spiculation	Degree of nodule spiculation	1. No Spiculation 2. Nearly No Spiculation 3. Medium Spiculation 4. Near Marked Spiculation 5. Marked Spiculation
Texture	Radiographic solidity of the nodule	1. Non-Solid/GGO 2. Non-Solid/Mixed 3. Part Solid/Mixed 4. Solid/Mixed 5. Solid
Malignancy	Subjective assessment of the likelihood of malignancy	1. Highly Unlikely 2. Moderately Unlikely 3. Indeterminate 4. Moderately Likely 5. Highly Likely

Abbreviations: GGO = Ground-glass opacity.

.2 Fixmatch pipeline

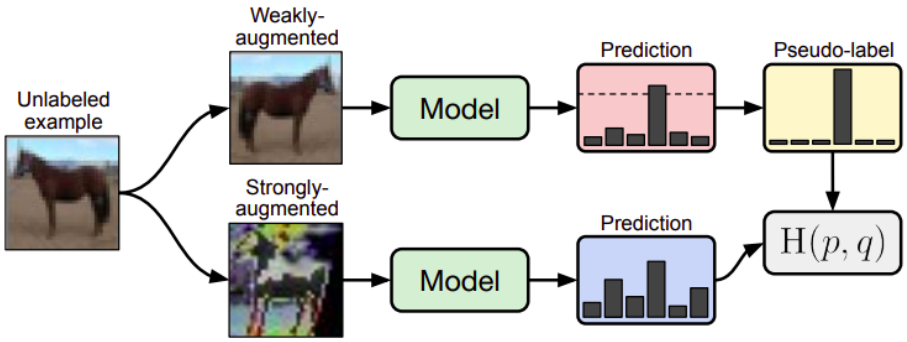


FIGURE A.1: Original FixMatch framework illustrating augmentation consistency and pseudo-labeling. Adapted from [14].

.3 LIDC-IDRI statistics

TABLE A.2: Inter-radiologist disagreement statistics for malignancy annotations in LIDC-IDRI.

Metric	Median = 3	Median $\neq 3$
Mean variance	0.3198	0.3554
Mean entropy	0.4590	0.4227
Proportion (%)	42.25%	57.75%

* *Note: Malignancy is an ordinal scale from 1 to 5, meaning that Median* $\neq 3$ *corresponds to cases* $\{1, 2, 4, 5\}$.

TABLE A.3: Inter-radiologist disagreement across Lung Nodule Visual Attributes, measured using variance and entropy over categorical annotations.

Attribute	Mean Variance	Mean Entropy
Sphericity	0.3109	0.4665
Subtlety	0.3807	0.4508
Margin	0.3741	0.4324
Lobulation	0.3309	0.3701
Spiculation	0.2882	0.3109
Texture	0.2393	0.2278
Calcification	0.1227	0.0534
Internal Structure	0.0131	0.0060

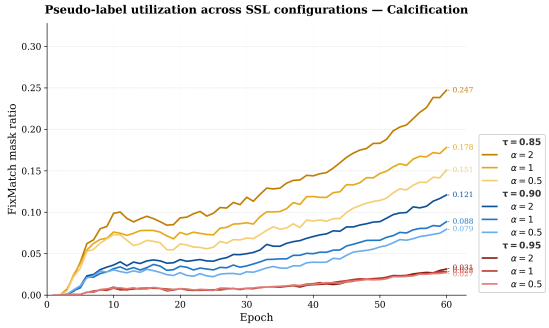
TABLE A.4: Class imbalance across LIDC-IDRI attributes.

Attribute	Majority Class (%)	Minority Class (%)	Imbalance Ratio
Nodule Internal Structure	1 (99.47%)	5 (0.04%)	2486.0
Nodule Calcification	6 (87.50%)	2 (0.08%)	1093.0
Nodule Sphericity	4 (40.27%)	1 (0.19%)	212.0
Nodule Lobulation	1 (69.22%)	5 (1.33%)	52.0
Nodule Spiculation	1 (76.50%)	5 (1.64%)	46.6
Nodule Texture	5 (68.84%)	2 (2.63%)	26.2
Nodule Malignancy	3 (42.25%)	5 (4.38%)	9.65
Nodule Margin	4 (35.43%)	1 (5.68%)	6.2
Nodule Subtlety	4 (31.66%)	1 (6.06%)	5.2

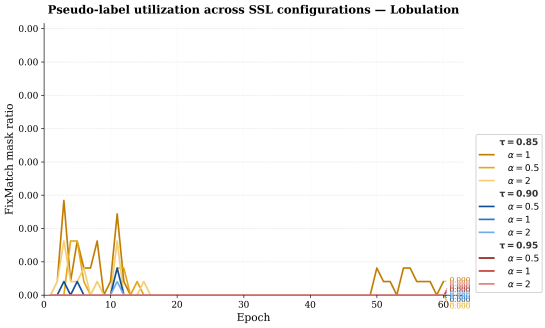
TABLE A.5: Class distribution of malignancy labels in LUNA25 dataset.

Attribute	Class	Count	Percentage
Nodule Malignancy	0 (Benign)	5608	90.99%
Nodule Malignancy	1 (Malignant)	555	9.01%

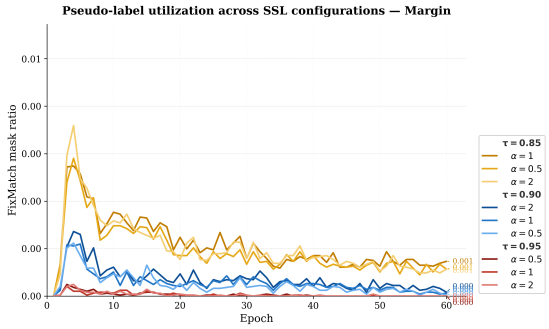
.4 SSL extra plots



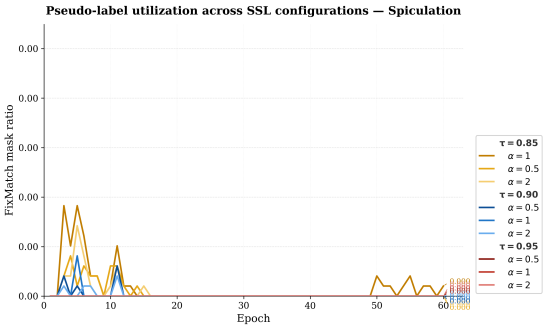
(A) Calcification



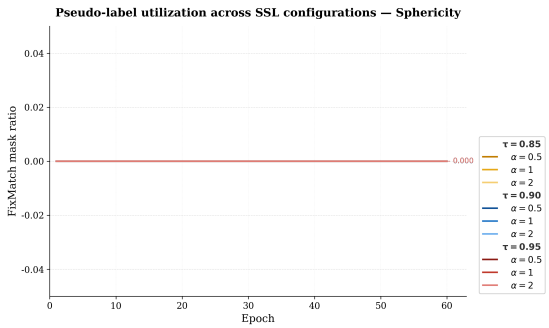
(B) Lobulation



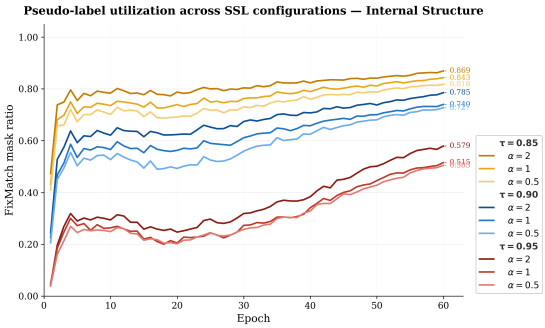
(C) Margin



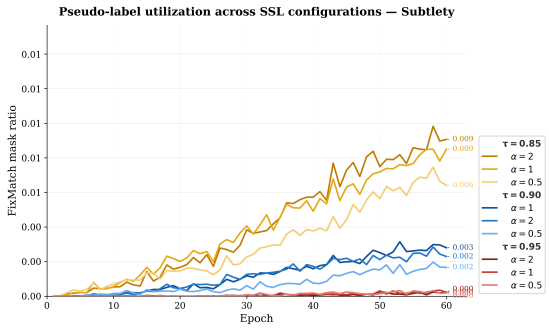
(D) Spiculation



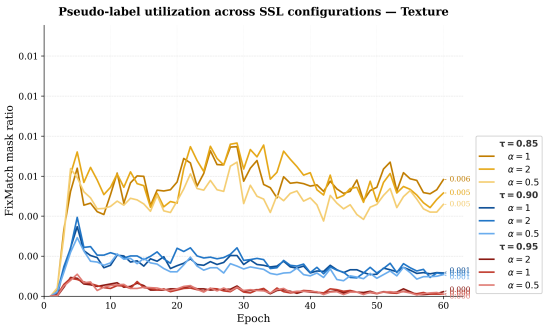
(E) Sphericity



(F) Structure



(G) Subtlety



(H) Texture

FIGURE A.2: FixMatch mask ratio across LNVA attributes.

Bibliography

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," vol. 74, no. 3, pp. 229–263. [Online]. Available: https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21834 [Cited on page 1.]

[2] S. G. Armato, M. F. McNitt-Gray, A. P. Reeves, C. R. Meyer, G. McLennan, D. R. Aberle, E. A. Kazerooni, H. MacMahon, E. J. Van Beek, D. Yankelevitz, E. A. Hoffman, C. I. Henschke, R. Y. Roberts, M. S. Brown, R. M. Engelmann, R. C. Pais, C. W. Piker, D. Qing, M. Kocherginsky, B. Y. Croft, and L. P. Clarke, "The lung image database consortium (LIDC): An evaluation of radiologist variability in the identification of lung nodules on CT scans," vol. 14, no. 11, pp. 1409–1421. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1076633207003984 [Cited on page 1.]

[3] M. P. Rivera, A. C. Mehta, and M. M. Wahidi, "Establishing the diagnosis of lung cancer," vol. 143, no. 5, pp. e142S–e165S. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0012369213602937 [Cited on page 1.]

[4] H. Sugimoto, Y. Fukuda, Y. Yamada, H. Ito, T. Tanaka, T. Yoshida, S. Okamori, K. Ando, and Y. Okada, "Complications of a lung biopsy for severe respiratory failure: A systematic review and meta-analysis," vol. 61, no. 1, pp. 121–132. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2212534522001204 [Cited on page 1.]

[5] R. Jiao, Y. Zhang, L. Ding, R. Cai, and J. Zhang, "Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation." [Online]. Available: http://arxiv.org/abs/2207.14191 [Cited on pages 1 and 5.]

[6] S. G. Armato, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. R. Van Beeke, D. Yankelevitz, A. M. Biancardi, P. H. Bland, M. S. Brown, R. M. Engelmann, G. E. Laderach, D. Max, R. C. Pais, D. P. Y. Qing, R. Y. Roberts, A. R. Smith, A. Starkey, P. Batrah, P. Caligiuri, A. Farooqi, G. W. Gladish, C. M. Jude, R. F. Munden, I. Petkovska, L. E. Quint, L. H. Schwartz, B. Sundaram, L. E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. V. Casteele, S. Gupte, M. Sallamm, M. D. Heath, M. H. Kuhn, E. Dharaiya, R. Burns, D. S. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, and B. Y. Croft, "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans," vol. 38, no. 2, pp. 915–931. [Cited on pages 1 and 10.]

[7] L. M. Pehrson, M. B. Nielsen, and C. Ammitzbøl Lauridsen, "Automatic pulmonary nodule detection applying deep learning or machine learning algorithms to the LIDC-IDRI database: A systematic review," vol. 9, no. 1, p. 29. [Online]. Available: https://www.mdpi.com/2075-4418/9/1/29 [Cited on pages 2 and 5.]

[8] S. Shen, S. X. Han, D. R. Aberle, A. A. Bui, and W. Hsu, "An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification," vol. 128, pp. 84–95. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417419300545 [Cited on pages 5 and 7.]

[9] X. Fu, L. Bi, A. Kumar, M. Fulham, and J. Kim, "Attention-enhanced cross-task network for analysing multiple attributes of lung nodules in CT," version Number: 2. [Online]. Available: https://arxiv.org/abs/2103.03931 [Cited on page 5.]

[10] M. Gouveia, J. Araújo, H. P. Oliveira, and T. Pereira, "Domain-specific data augmentation for lung nodule malignancy classification," in *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/11254063/ [Cited on page 5.]

[11] J. Pedrosa, G. Aresta, C. Ferreira, M. Rodrigues, P. Leitão, A. S. Carvalho, J. Rebelo, E. Negrão, I. Ramos, A. Cunha, and A. Campilho, "LNDb: A lung nodule database on computed tomography," version Number: 3. [Online]. Available: https://arxiv.org/abs/1911.08434 [Cited on page 5.]

[12] S. G. Armato, K. Drukker, F. Li, L. Hadjiiski, G. D. Tourassi, R. M. Engelmann, M. L. Giger, G. Redmond, K. Farahani, J. S. Kirby, and L. P. Clarke, "LUNGx challenge for computerized lung nodule classification," vol. 3, no. 4, p. 044506. [Cited on page 5.]

[13] I. D. Apostolopoulos, N. D. Papathanasiou, and G. S. Panayiotakis, "Classification of lung nodule malignancy in computed tomography imaging utilising generative adversarial networks and semi-supervised transfer learning," vol. 41, no. 4, pp. 1243–1257. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0208521621000991 [Cited on pages 6 and 7.]

[14] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," version Number: 2. [Online]. Available: https://arxiv.org/abs/2001.07685 [Cited on pages 6, 7, 14, and 70.]

[15] A. J. P. O. E. Abreu, "Overcoming scarce annotations through deep semi-supervised learning in cancer characterisation," accepted: 2026-01-26T01:29:44Z. [Online]. Available: https://repositorio-aberto.up.pt/handle/10216/169698 [Cited on pages 6 and 7.]

[16] N. Khosravan and U. Bagci, "Semi-supervised multi-task learning for lung cancer diagnosis," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, pp. 710–713. [Online]. Available: https://ieeexplore.ieee.org/document/8512294/ [Cited on pages 6 and 7.]

[17] W. Zhu, C. Liu, W. Fan, and X. Xie, "DeepLung: Deep 3d dual path nets for automated pulmonary nodule detection and classification," version Number: 1. [Online]. Available: https://arxiv.org/abs/1801.09555 [Cited on pages 6 and 7.]

[18] F. I. Tushar, A. Wang, L. Dahal, E. Samei, M. R. Harowicz, J. Kalpathy-Cramer, K. J. Lafata, T. D. Tailor, C. Rudin, and J. Y. Lo, "Reproducible benchmarking for lung nodule detection and malignancy classification across multiple low-dose CT datasets." [Online]. Available: http://arxiv.org/abs/2405.04605 [Cited on pages 6 and 11.]

[19] M. C. Hancock and J. F. Magnan, "Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: probing the lung image database consortium dataset with two statistical learning methods," vol. 3, no. 4, p. 044504. [Online]. Available: http://medicalimaging.spiedigitallibrary.org/article.aspx?doi=10.1117/1.JMI.3.4.044504 [Cited on pages 7, 40, and 61.]

[20] S. Hussein, K. Cao, Q. Song, and U. Bagci, "Risk stratification of lung nodules using 3d CNN-based multi-task learning," version Number: 1. [Online]. Available: https://arxiv.org/abs/1704.08797 [Cited on page 7.]

[21] R. LaLonde, D. Torigian, and U. Bagci, "Encoding visual attributes in capsules for explainable medical diagnoses," version Number: 5. [Online]. Available: https://arxiv.org/abs/1909.05926 [Cited on page 8.]

[22] E. M. Rodrigues, M. Gouveia, H. P. Oliveira, and T. Pereira, "Efficient-proto-caps: A parameter-efficient and interpretable capsule network for lung nodule characterization," vol. 13, pp. 56 616–56 630. [Online]. Available: https://ieeexplore.ieee.org/document/10943124/ [Cited on pages 8, 11, 13, and 68.]

[23] H. Zhang, X. Gu, M. Zhang, W. Yu, L. Chen, Z. Wang, F. Yao, Y. Gu, and G.-Z. Yang, "Re-thinking and re-labeling LIDC-IDRI for robust pulmonary cancer prediction." [Online]. Available: http://arxiv.org/abs/2207.14238 [Cited on page 10.]

[24] The National Lung Screening Trial Research Team, "Reduced lung-cancer mortality with low-dose computed tomographic screening," vol. 365, no. 5, pp. 395–409. [Online]. Available: http://www.nejm.org/doi/10.1056/NEJMoa1102873 [Cited on page 11.]

[25] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," version Number: 2. [Online]. Available: https://arxiv.org/abs/2208.09910 [Cited on pages 15 and 16.]