Optimized Pruning Strategies for Non-local Architectures in Lung Computed Tomography Tumor Recognition

Aleksei Samarin, Aleksei Toropov, Artem Nazarenko
ITMO University / ISP RAS
St. Petersburg, Russia
avsamarin@itmo.ru, toropov.ag@hotmail.com, aanazarenko@itmo.ru

Egor Kotenko
St. Petersburg State University / ISP RAS
St. Petersburg, Russia
kotenkoed@gmail.com

Anastasia Mamaeva, Dmitry Nazarenko
ITMO University
St. Petersburg, Russia
asmamaeva@itmo.ru, nazarenkodmit@gmail.com

Valentin Malykh
International IT University
Almaty, Kazakhstan
valentin.malykh@phystech.edu

Elena Mikhailova
St. Petersburg State University
St. Petersburg, Russia
e.mikhailova@itmo.ru

Alexander Savelev
St. Petersburg Electrotechnical University "LETI" / ISP RAS
St. Petersburg, Russia
algsavelev@gmail.com

Alexander Motyko
St. Petersburg Electrotechnical University "LETI"
St. Petersburg, Russia
aamotyko@etu.ru

***Abstract*—Attention-augmented encoder-decoder models with non-local modules are effective for lung tumor analysis on computed tomography (CT), yet their computational and memory demands can hinder deployment. We study pruning for a specialized non-local U-Net pipeline that jointly performs neoplasm presence recognition and lesion segmentation from single-channel CT snapshots. We benchmark unstructured and structured pruning baselines and propose a sensitivity-aware, architecture-guided refinement that allocates sparsity across encoder-decoder components and attention modules to better preserve task-critical capacity. On a joint benchmark of public lung CT datasets, the proposed method achieves a superior accuracy-efficiency trade-off. At 50% sparsity, it reduces latency from 41.5 ms to 23.6 ms and FLOPs from 62.4G to 30.8G while maintaining strong quality (F1 = 0.923, mIoU = 0.742, Dice = 0.847). At 60% sparsity, it further reduces latency to 19.4 ms (FLOPs = 23.5G) with stable performance (F1 = 0.904, mIoU = 0.724). Overall, the proposed pruning refinement improves the deployability of attention-based lung CT models without sacrificing clinically meaningful performance.**

***Index Terms*—lung computed tomography, tumor segmentation, neoplasm classification, non-local attention, model pruning**

I. Introduction

Deep neural networks have steadily evolved toward higher representational capacity, but this progress has come with a tangible cost in compute, memory footprint, and engineering complexity. Architectural trends that improved accuracy, including deeper residual backbones, compound scaling, and attention-centric designs, often increase parameter count, activation memory, and inference latency, especially at clinically relevant image resolutions [1]–[4]. Importantly, the burden is not limited to large foundation models. Even task-specific vision pipelines become expensive once they incorporate multi-branch heads, multi-task objectives, or attention modules whose complexity scales unfavorably with spatial extent [5]. For this reason, resource constraints increasingly shape what can be deployed, not only what can be trained.

These constraints are particularly acute in medicine and biomedical imaging, where computational budgets and latency requirements are tightly coupled to clinical utility. Unlike many natural-image settings, medical imaging pipelines often process high-resolution studies and, in many cases, multiple slices per examination, which amplifies both activation memory and end-to-end inference time. In real deployments, the same model may be expected to operate across heterogeneous environments that range from centralized servers to standard hospital workstations and edge devices, while still meeting throughput and turnaround constraints imposed by clinical workflows. Furthermore, biomedical systems frequently require more than a single prediction. Segmentation, lesion localization, and study-level classification are commonly coupled within the same pipeline, and the overall reliability depends on maintaining consistent performance across scanners, protocols, and institutions [6], [7]. At the same time, the operational value of deep learning in healthcare is directly linked to scalability. Screening, triage, and decision support benefit from fast inference, predictable latency, and lower computational overhead, which in turn reduces infrastructure costs and broadens access to AI-assisted analysis [8], [9]. This deployment gap has motivated a growing body of work on lightweight or deployment-oriented medical models, including

recent efficient pipelines for infrared eye structure segmentation [10], [11].

Model compression offers several complementary routes, including quantization, distillation, low-rank factorization, and pruning. Among them, pruning remains one of the most direct and widely applicable strategies when the goal is to reduce model size and accelerate inference with minimal architectural disruption [12]. In general, pruning methods range from unstructured weight removal, often magnitude-based, to structured removal of channels, filters, or blocks that map more reliably to real hardware speedups [13]–[16]. However, a recurring lesson from the pruning literature is that applying standard baselines without adaptation rarely yields the best accuracy-efficiency trade-off across architectures and tasks. Layer sensitivity is strongly non-uniform, attention modules can respond differently than convolutional blocks, and the optimal pruning schedule is often intertwined with fine-tuning dynamics [17], [18]. These issues are particularly pronounced in multi-task medical pipelines, where naive sparsification can degrade segmentation and classification asymmetrically.

Motivated by these considerations, we investigate pruning in the setting of our previously proposed lung computer tomography pipeline for joint neoplasm segmentation and presence recognition from single-channel computer tomography snapshots [19]. The baseline model is an attention-augmented U-Net that integrates specialized non-local blocks and achieves strong performance on lung tumor analysis [5], [6]. Building on this architecture, we perform a systematic comparison of multiple pruning strategies and evaluate their impact on predictive quality and computational footprint. Furthermore, we propose an optimization of the pruning procedure tailored to the structural properties of the studied attention modules and to the constraints of biomedical deployment. Overall, our study provides an empirical and methodological basis for obtaining more resource-efficient specialized attention models for lung computer tomography analysis, while maintaining clinically meaningful accuracy. Our experiments further show that parameter sparsity alone does not guarantee practical acceleration in this setting: unstructured pruning yields limited wall-clock improvements despite substantial parameter reduction. In contrast, the proposed sensitivity-aware strategy achieves a consistently better accuracy-efficiency trade-off at moderate-to-high sparsity levels, providing substantial latency and FLOPs reductions while preserving both segmentation fidelity and presence recognition quality. We also explore efficiency-oriented model design in related vision tasks beyond lung CT analysis, including lightweight image enhancement and color correction pipelines and their training variants [20], [21].

II. Related work

Early deep learning systems for biomedical image analysis primarily relied on fully convolutional encoder-decoder designs. In particular, U-Net became a widely adopted baseline due to its strong inductive bias for dense prediction and its ability to aggregate multi-scale context via skip connections [6]. Over time, this family of architectures was refined and applied across many clinical tasks, and a number of best practices for medical imaging pipelines were consolidated in broad surveys [7]. Despite their empirical success, purely convolutional models have a well-known limitation. Their receptive field grows indirectly through depth and pooling, so long-range dependencies and global context may be represented inefficiently. In lung CT analysis, where subtle lesions can be confounded by surrounding anatomy and acquisition variability, this limitation can translate into reduced robustness or a need for larger and slower models, which complicates deployment.

Subsequently, the community began integrating explicit global interaction mechanisms into vision architectures to overcome the locality bias of convolutions. Non-local neural networks formalized the idea of computing a response at a position as a weighted aggregation over all positions, enabling direct long-range feature interactions [5]. In principle, non-local modeling is attractive for medical imaging because it can capture context beyond local neighborhoods, which is often helpful when the evidence for pathology is diffuse or when discriminative cues are spatially separated. However, this benefit comes with practical drawbacks. Non-local and attention-like operations increase computational and memory costs, especially when applied to high-resolution feature maps that are common in biomedical segmentation. This is particularly problematic in clinical pipelines that require more than a single output, since segmentation, lesion localization, and study-level classification are often coupled, and the associated decoders and heads amplify activation memory and end-to-end latency. A line of recent work has therefore explored task-specific attention adaptations that aim to retain the benefits of global context while maintaining feasible inference cost [19].

More recently, attention mechanisms were further popularized by the transformer architecture [3] and later adapted to vision at scale [4]. Transformer-style models strengthened the case for global interaction modeling, but they also highlighted a persistent challenge for biomedical imaging. The core attention operation can scale unfavorably with spatial resolution, which can be prohibitive for dense prediction tasks unless additional approximations are introduced. In medical contexts, this often leads to higher deployment cost, increased engineering complexity, or more restrictive input preprocessing. As a result, even when attention improves accuracy, practicality can lag, reinforcing the gap between research-grade performance and clinically usable solutions. Related deployment-oriented studies in other biomedical imaging modalities also report that attention-enhanced designs require careful efficiency considerations to remain practical [10], [11].

In parallel with architectural advances, model compression developed as a practical response to rising inference costs. Among compression techniques, pruning has remained one of the most widely used approaches because it can be applied to already trained networks and can reduce parameter count and, in structured variants, inference latency [12]. Early and influential pruning recipes include magnitude-based removal

of weights or units, as well as structured pruning of channels and filters that align better with efficient execution on commodity hardware [14], [15]. Shortly thereafter, gradient- and sensitivity-driven criteria were introduced to better estimate the impact of removing parameters, enabling more informed pruning decisions than purely magnitude-based heuristics [16]. Nevertheless, baseline pruning strategies exhibit an important limitation. Pruning sensitivity is highly non-uniform across layers, so applying a global sparsity target can remove capacity from critical layers or modules, producing disproportionate accuracy loss. This effect can be amplified in architectures that include attention or non-local components, where representation bottlenecks and compute distribution differ from standard convolutional blocks.

In subsequent large-scale analyses, another practical issue became evident. The relationship between sparsity and real-world speedup is not guaranteed. Unstructured sparsity often fails to translate into wall-clock gains without specialized kernels, while structured sparsity is more likely to yield measurable acceleration but can be harder to tune without harming accuracy [13]. Around the same time, training dynamics were recognized as a first-order factor in successful sparsification. The lottery ticket hypothesis suggested that sparse subnetworks can train effectively when identified appropriately, implying that sparsity patterns and optimization interact in non-trivial ways [17]. Building on this intuition, movement pruning demonstrated that selecting which weights to prune during fine-tuning can outperform static, post hoc pruning criteria [18]. While these methods provide strong general tools, they are typically not tailored to specialized attention-augmented medical architectures, and they do not directly address multi-objective biomedical settings where segmentation fidelity and study-level classification must be preserved simultaneously.

Taken together, prior work suggests a consistent picture. Convolutional encoder-decoder baselines remain effective but may underutilize global context [6], [7]. Non-local and transformer-style attention improves long-range modeling but often increases computational and memory costs in high-resolution biomedical pipelines [3]–[5]. Pruning offers a practical path toward efficiency, yet generic pruning baselines can be suboptimal for attention-augmented, multi-task medical models due to layer-dependent sensitivity, limited transferability of pruning recipes across module types, and the coupling between pruning schedules and fine-tuning dynamics [13]–[18]. These gaps motivate our study of pruning strategies for a specialized non-local U-Net style lung CT model, including an additional optimization of the pruning procedure designed to improve the accuracy-efficiency trade-off under biomedical deployment constraints.

III. Problem statement

Attention-augmented encoder-decoder architectures with specialized non-local modules achieve strong performance for lung CT neoplasm segmentation and presence recognition, yet their computational and memory demands can limit practical deployment under constrained clinical hardware and latency budgets. Pruning is a natural approach to reduce model size and inference cost, but off-the-shelf pruning baselines are frequently suboptimal for models that combine convolutional backbones with attention components and multi-task objectives. In such settings, pruning sensitivity varies substantially across layers and modules, and sparsification can affect segmentation and classification unevenly, leading to an unfavorable accuracy-efficiency trade-off.

This work aims to derive a resource-efficient variant of a specialized non-local U-Net style lung CT model through pruning, while preserving clinically meaningful quality for both segmentation and classification. To this end, we systematically benchmark representative pruning strategies, quantify their impact on predictive performance and computational footprint relative to the unpruned baseline, and propose an architecture-aware refinement of the pruning procedure intended to improve the resulting accuracy-efficiency frontier compared to standard pruning recipes.

IV. Proposed solution

Pruning methods reduce the computational and memory footprint of a neural network by removing parameters or structural components while attempting to preserve predictive quality. Formally, given a network $f(x;\theta)$ with parameters $\theta \in \mathbb{R}^d$ and a training objective

$$\min_{\theta} \ \mathcal{L}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell(f(x;\theta), y)\right], \tag{1}$$

pruning can be expressed via a binary mask $m \in \{0,1\}^d$ applied to parameters, yielding an effective parameterization $\theta \odot m$. A generic pruning objective seeks a sparse mask under a resource constraint:

$$\min_{\theta,m} \ \mathcal{L}(\theta \odot m) \quad \text{s.t.} \quad \|m\|_0 \le s, \tag{2}$$

where s controls the number of retained parameters. In practice, this problem is typically approximated by heuristic criteria and a fine-tuning stage.

Historically, pruning originated from second-order saliency criteria that estimate the loss increase caused by removing a parameter, as introduced in Optimal Brain Damage and extended in Optimal Brain Surgeon [22], [23]. While these formulations provide a principled perspective, their direct application to modern deep networks is often impractical due to the computational cost of curvature estimation, which motivated the widespread use of simpler criteria.

A dominant family of baselines relies on magnitude-driven unstructured pruning, where individual weights are removed based on their absolute values, typically followed by fine-tuning [12], [13], [24]. A common instantiation is global magnitude pruning with a threshold τ, where the mask is defined as

$$m_i = \mathbb{I}\left(|\theta_i| \ge \tau\right), \tag{3}$$

and τ is chosen to satisfy the target sparsity in (2). Variants that incorporate pruning schedules, such as gradual sparsification during training, can improve stability compared to one-shot pruning and reduce the need for extensive manual tuning [24].

Dynamic approaches that allow connections to be removed and, if needed, reintroduced during training further emphasize that pruning decisions may be error-prone and benefit from corrective mechanisms [25]. Despite their simplicity and strong compression ratios, unstructured sparsity does not reliably translate into wall-clock speedups on standard hardware without dedicated sparse kernels and runtime support [13].

Structured pruning provides a more deployment-oriented alternative by removing hardware-friendly units such as channels, filters, or blocks. Let $\{\mathcal{G}_k\}_{k=1}^{K}$ denote a partition of parameters into groups corresponding to such structures (for example, all weights of one convolutional channel). A standard way to encourage structured sparsity is to add a group regularizer

$$\min_{\theta} \ \mathcal{L}(\theta) + \lambda \sum_{k=1}^{K} \|\theta_{\mathcal{G}_k}\|_2, \tag{4}$$

where groups with small ℓ_2 norms can be removed to obtain a compact architecture. Early and influential methods include filter pruning based on heuristic importance measures [14], channel selection via sparsity-inducing regularization as in network slimming [15], and sensitivity-based criteria derived from gradient or Taylor approximations [16]. A typical Taylor-style importance score for a structured unit u is

$$I(u) \approx \left| \frac{\partial \mathcal{L}}{\partial a_u} a_u \right|, \tag{5}$$

where a_u denotes the activation (or output) associated with u, and units with low $I(u)$ are pruned. Later, structured approaches were extended to more explicit optimization formulations and reconstruction-based criteria, including channel pruning via LASSO-style selection and feature reconstruction [26], as well as filter-level pruning frameworks such as ThiNet [27]. Regularization-based approaches that directly promote structured sparsity during training further support pruning of channels, filters, and even layer structures in a way that is compatible with efficient execution [28]. In addition, compression methods related to pruning, such as low-rank tensor decompositions for convolutional layers, can reduce compute while preserving accuracy through discriminative fine-tuning [29].

Another key axis is the pruning schedule and the role of training dynamics. Beyond the classical train-prune-finetune pipeline, a line of work has questioned whether inherited weights are necessary and highlighted the importance of fair baselines, including training the target compact architecture from scratch [30]. Initialization-time pruning methods aim to reduce the dependence on expensive pruning schedules, for example by selecting connections using sensitivity at initialization [31]. Data-agnostic criteria were proposed to mitigate failure modes such as layer collapse when pruning at initialization [32]. Sparse training methods that update sparse connectivity patterns during training, such as RigL, suggest that sparse topology and optimization should be treated jointly rather than sequentially [33].

Finally, pruning and compression can be guided by direct deployment metrics. Platform-aware adaptation methods incorporate measured latency and energy into the simplification loop, which is often more informative than indirect proxies such as parameter count or FLOPs [34]. Automated policy learning for compression, for example with reinforcement learning, further demonstrates that compression choices can be optimized at the system level [35]. Structure learning methods such as MorphNet iteratively reshape networks under explicit resource constraints and can be viewed as complementary to pruning when the goal is to satisfy a compute budget [36]. In attention-based models, additional pruning dimensions appear, such as pruning attention heads, where empirical evidence indicates that many heads can be removed with limited impact on accuracy [37]. Overall, these baselines cover a wide spectrum of pruning and compression strategies, and they highlight that the best accuracy-efficiency trade-off is strongly architecture- and deployment-dependent, motivating the refinement proposed in this work.

A. Customize pruning strategies

The goal of the proposed strategy is to improve the accuracy-efficiency trade-off of pruning for an attention-augmented U-Net style architecture with specialized non-local modules in a joint segmentation and classification setting. The main motivation is that generic baselines typically distribute sparsity in a way that is not aligned with the computational profile and sensitivity of attention and encoder-decoder components. As a result, naive pruning can either over-prune modules that are critical for segmentation fidelity or preserve parameters in regions that contribute little to quality while dominating inference cost.

B. Customized pruning strategy for lung CT analysis

While standard pruning baselines can reduce model size and computational cost, their application to attention-augmented medical architectures often yields suboptimal accuracy-efficiency trade-offs. As discussed in the problem statement, pruning sensitivity varies considerably across encoder-decoder stages and attention modules, and naive sparsity allocation can disproportionately remove capacity from components essential for preserving fine-grained lesion boundaries and stable global context modeling. To address these limitations, we propose a customized pruning strategy designed specifically for the non-local U-Net style pipeline in the lung CT setting. The proposed approach incorporates two key refinements: (i) a sensitivity-aware sparsity allocation mechanism that accounts for the functional role of different modules, and (ii) a targeted retention of latent representations corresponding to diagnostically relevant anatomical regions. As illustrated in Fig. 1, the proposed strategy follows a two-stage procedure: we first estimate encoder redundancy from dataset-level feature statistics and then use the resulting statistics to guide an architecture-aware reduction of encoder computations during pruning and fine-tuning.

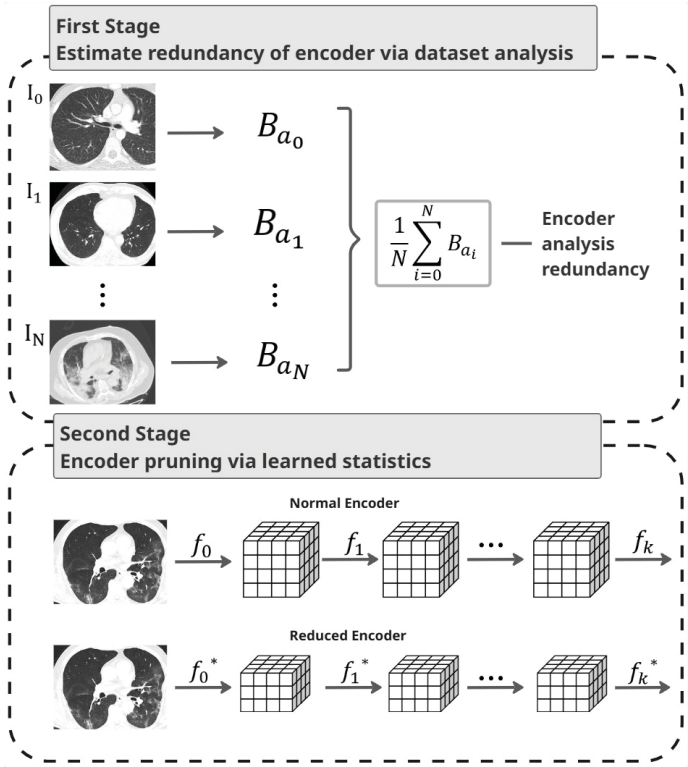

Fig. 1. Two-stage customized pruning pipeline. Stage 1 estimates encoder redundancy from dataset-level feature statistics. Stage 2 uses the learned statistics to derive a reduced encoder by replacing operations $\{f_k\}$ with their pruned variants $\{f_k^*\}$.

1) Activation norm-guided importance scoring: Standard magnitude-based pruning criteria operate solely on weight values and fail to capture the actual contribution of parameters to the forward pass. In medical imaging, where background regions dominate the input space and foreground lesions occupy only a small fraction of the field-of-view, this limitation is particularly consequential. Parameters that contribute primarily to background processing may receive similar magnitude scores as those critical for lesion detection, leading to indiscriminate pruning that degrades diagnostic performance.

To address this, we incorporate activation statistics into the importance assessment. Let $a_u \in \mathbb{R}^{H\times W}$ denote the spatial activation map produced by a structured unit u (e.g., a convolutional channel or attention head) for a given input. For each unit, we compute an activation-weighted importance score:

$$I_{\text{act}}(u) = \frac{1}{N}\sum_{n=1}^{N} \left\| a_u^{(n)} \right\|_1, \tag{6}$$

where N is the number of samples in a calibration set drawn from the training distribution. This score reflects the average magnitude of activations produced by the unit across the dataset. Units that consistently produce weak activations contribute minimally to downstream representations and are candidates for removal. However, this global average can still be dominated by background regions. To refine the score, we introduce a foreground-aware variant:

$$I_{\text{fg}}(u) = \frac{1}{N}\sum_{n=1}^{N} \left\| a_u^{(n)} \odot M_{\text{fg}}^{(n)} \right\|_1, \tag{7}$$

where $M_{\text{fg}}^{(n)}$ is a binary foreground mask indicating lesion regions (obtained from segmentation ground truth or derived from annotations). This formulation explicitly prioritizes units that contribute to processing diagnostically relevant anatomy.

2) Latent space background suppression: A distinctive characteristic of lung CT analysis is the prevalence of background regions corresponding to air-filled lung parenchyma and extrapulmonary space. These regions, while occupying substantial spatial extent in the input and intermediate feature maps, carry limited diagnostic information for tumor recognition and segmentation. In the latent space of the encoder-decoder architecture, background regions manifest as activation patterns that are consistently low-magnitude and exhibit minimal variation across inputs. Crucially, processing these regions consumes computational resources without contributing meaningfully to the final predictions.

We exploit this observation by introducing a structured pruning mechanism that operates at the level of spatial positions in the latent feature maps. Specifically, we identify a set of spatial locations that correspond to background regions with high confidence. Let $F \in \mathbb{R}^{C\times H\times W}$ denote a feature map at a given depth in the encoder. For each spatial position (i, j), we compute a background confidence score:

$$B(i,j) = \sigma\left(\frac{\|F_{:,:,i,j}\|_2 - \mu_b}{\sigma_b}\right), \tag{8}$$

where μ_b and σ_b are the mean and standard deviation of background region activations estimated from the calibration set, and $\sigma(\cdot)$ denotes the sigmoid function. Positions with $B(i, j) > \tau_b$ are designated as high-confidence background, where τ_b is a threshold selected to achieve a desired level of spatial sparsity.

For these identified positions, we can safely omit computation in subsequent layers by applying a spatial pruning mask that zeros out the corresponding activations. This is equivalent to introducing a hard attention mechanism that suppresses background regions early in the encoder path. The mask is applied identically across all channels for a given spatial position:

$$\tilde{F}_{:,i,j} = \mathbb{I}\left(B(i,j) \le \tau_b\right) \cdot F_{:,i,j}. \tag{9}$$

This operation reduces the effective spatial resolution of feature maps processed by downstream layers, yielding proportional reductions in FLOPs and memory footprint for convolutional and attention operations that scale with spatial extent. Importantly, the background confidence estimation can be performed efficiently using a lightweight projection head trained jointly with the main architecture, or derived from the segmentation decoder's intermediate representations.

3) Sensitivity-guided layerwise sparsity allocation: To address the non-uniform sensitivity of different architectural components, we introduce a layerwise sparsity allocation scheme based on estimated contribution to task performance. Following the Taylor expansion perspective in Eq. (5), we compute an importance score for each layer ℓ as:

$$S_\ell = \sum_{u \in \mathcal{U}_\ell} \left| \frac{\partial \mathcal{L}}{\partial a_u} \odot a_u \right|_1, \tag{10}$$

where $\mathcal{U}_\ell$ denotes the set of structured units in layer ℓ, and the gradient is evaluated on the calibration set. Layers with higher S_ℓ are considered more critical and are assigned lower target sparsity, while layers with lower S_ℓ are pruned more aggressively. The target sparsity s_ℓ for layer ℓ is determined by solving:

$$\min_{\{s_\ell\}} \sum_\ell S_\ell \cdot s_\ell \quad \text{s.t.} \quad \frac{\sum_\ell p_\ell \cdot (1 - s_\ell)}{\sum_\ell p_\ell} = t, \tag{11}$$

where p_ℓ is the parameter count of layer ℓ, and t is the overall parameter retention target (e.g., 0.5 for 50% pruning). This linear programming formulation allocates sparsity budget preferentially to layers with lower sensitivity, preserving capacity where it matters most.

4) Pruning schedule and fine-tuning: The proposed pruning strategy is implemented within an iterative pruning-fine-tuning schedule. Starting from the pretrained baseline, we alternate between pruning steps (removing a small fraction of parameters or spatial positions based on the criteria above) and fine-tuning steps to recover accuracy. This gradual approach allows the model to adapt to the changing architecture and mitigates the risk of irreversible capacity loss.

For attention modules specifically, we apply a modified importance score that accounts for head-level contributions. Following observations that many attention heads can be pruned with limited impact [37], we compute head importance as:

$$I_{\text{head}}(h) = \mathbb{E}_{x \sim \mathcal{D}_{\text{cal}}} \left[\| \text{Attn}_h(x) \|_F \right], \tag{12}$$

where $\text{Attn}_h(x)$ denotes the output of attention head h for input x, and $\| \cdot \|_F$ is the Frobenius norm. Heads with consistently low output norms are pruned first.

5) Summary: The proposed customized pruning strategy integrates three complementary mechanisms: (i) activation norm-guided importance scoring that prioritizes units contributing to foreground processing, (ii) spatial background suppression that eliminates computation on high-confidence background regions in the latent space, and (iii) sensitivity-guided layerwise allocation that distributes sparsity non-uniformly based on estimated contribution to task performance. Together, these refinements are designed to preserve diagnostically relevant representations while aggressively reducing computational footprint in regions of the model that contribute minimally to the final predictions. The following section evaluates this strategy against standard pruning baselines on the joint lung CT benchmark.

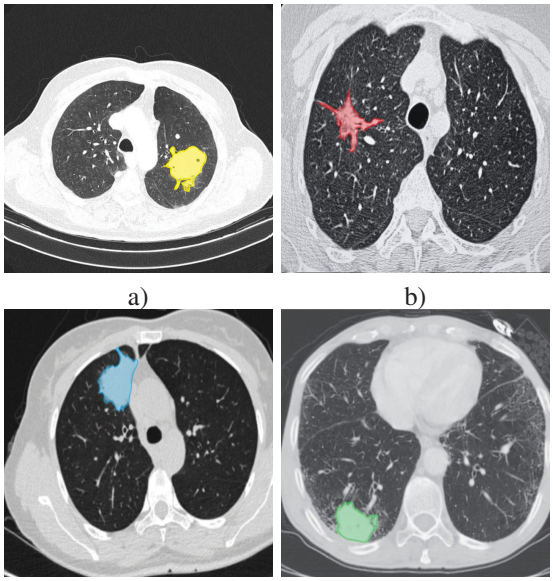

Fig. 2. Examples of lung CT snapshots from the joint benchmark illustrating different tumor types: (a) adenocarcinoma; (b) small-cell carcinoma; (c) large-cell carcinoma; (d) squamous cell carcinoma.

V. Experiments

A. Dataset description

To assess how pruning affects both segmentation and presence recognition, we benchmark baseline and customized pruning strategies on a joint lung CT dataset assembled from multiple publicly available sources. This setup allows us to measure accuracy-efficiency trade-offs under cross-source variability. The benchmark includes LIDC-IDRI, IQ-OTH/NCCD, Lung-PET-CT-Dx, and additional data obtained from Radiology Moscow and The Cancer Imaging Archive. Representative CT slices illustrating different tumor types are shown in Fig. 2.

1) LIDC-IDRI: LIDC-IDRI provides thoracic CT studies acquired for lung cancer screening and diagnosis, accompanied by lesion annotations produced by multiple radiologists. The dataset includes more than 1,000 cases and contains nodule-level information such as location and radiological characteristics, which makes it suitable for forming segmentation targets and study-level labels [38], [39].

2) IQ-OTH/NCCD: IQ-OTH/NCCD consists of CT scans from both healthy subjects and lung cancer patients spanning different disease stages. The dataset includes expert annotations from oncologists and radiologists and comprises 1,190 2D images extracted from 110 cases, enabling supervised learning for both presence recognition and lesion delineation [40].

3) Lung-PET-CT-Dx: Lung-PET-CT-Dx contains paired PET and CT examinations of lung cancer patients with corresponding tumor-related annotations. Although the dataset is multimodal, in this work we use the CT component and leverage the provided tumor information for constructing consistent targets within our unified evaluation pipeline. The

dataset contains more than 200,000 slices from 355 patients [41].

4) Radiology Moscow and the cancer imaging archive: In addition to the above sources, we use a combined subset of CT slice data obtained from Radiology Moscow and The Cancer Imaging Archive. This subset contains 10,052 single-channel images and is used for training, validation, and testing under a balanced binary labeling scheme. The data are partitioned into 7,196 images for training, 1,428 for validation, and 1,428 for testing, with equal representation of neoplasm-present and neoplasm-absent classes [42].

For a portion of the collected sources, the original lesion markup is provided in a detection-oriented format. To enable a consistent segmentation evaluation across datasets, we convert such markup into binary segmentation masks using a watershed-based procedure applied to the region of interest. Binary labels for presence recognition are derived from the corresponding metadata and the provided clinical annotations.

B. Experimental results

We evaluate the impact of pruning on the proposed lung CT pipeline in the joint setting of neoplasm presence recognition and lesion segmentation. All pruning strategies are applied to the same baseline architecture, namely the attention-augmented U-Net with specialized S-L non-local blocks (denoted as Base). This design choice enables a direct comparison of pruning methods by isolating the effect of sparsification from architectural differences.

For the classification task, the benchmark is organized into two classes, neoplasm-present and neoplasm-absent, and we report Precision, Recall, and F1-score. For the segmentation task, we use binary lesion masks and report mean Intersection-over-Union (mIoU) and Dice coefficient (mDice). In addition to predictive quality, we quantify the computational footprint of each pruned variant by reporting the number of trainable parameters, FLOPs, and inference latency measured under a fixed hardware configuration and identical preprocessing. Together, these measurements characterize the accuracy-efficiency trade-off induced by different pruning strategies and enable a fair comparison against the unpruned baseline.

Table I summarizes classification performance, Table II reports segmentation quality, and Table III provides efficiency metrics. The tables are intentionally left with placeholders and will be populated after completing the full experimental sweep across baseline pruning strategies and the proposed customized pruning procedure.

The resulting evaluation protocol is intended to highlight how different pruning baselines and the proposed customized strategy affect the trade-off between predictive quality and deployment-relevant efficiency when applied to the same attention-augmented lung CT model.

TABLE I. CLASSIFICATION RESULTS UNDER PRUNING. BASE DENOTES THE UNPRUNED S-L NLB U-NET.

Method	Sparsity	Precision	Recall	F1
Base (Unpruned)	0%	0.942	0.931	0.936
GMP	30%	0.931	0.914	0.922
GMP	50%	0.901	0.879	0.890
Structured Channel	30%	0.937	0.924	0.930
Structured Channel	50%	0.917	0.896	0.906
Structured Channel	60%	0.891	0.868	0.879
Attention Head Pruning	40%	0.923	0.905	0.914
Attention Head Pruning	60%	0.888	0.861	0.874
Proposed Sensitivity-Aware	30%	0.939	0.926	0.932
Proposed Sensitivity-Aware	50%	0.932	0.915	0.923
Proposed Sensitivity-Aware	60%	0.911	0.897	0.904

TABLE II. SEGMENTATION RESULTS UNDER PRUNING. BASE DENOTES THE UNPRUNED S-L NLB U-NET.

Method	Sparsity	mIoU	mDice
Base (Unpruned)	0%	0.781	0.872
GMP	30%	0.754	0.848
GMP	50%	0.706	0.817
Structured Channel	30%	0.768	0.859
Structured Channel	50%	0.734	0.842
Structured Channel	60%	0.712	0.826
Attention Head Pruning	40%	0.746	0.845
Attention Head Pruning	60%	0.718	0.831
Proposed Sensitivity-Aware	30%	0.772	0.864
Proposed Sensitivity-Aware	50%	0.742	0.847
Proposed Sensitivity-Aware	60%	0.724	0.835

VI. CONCLUSION

In this work, we studied pruning as a practical approach to improve the deployability of an attention-augmented U-Net with specialized non-local blocks for lung tumor analysis from single-channel CT snapshots. The evaluation was conducted in a joint setting that includes neoplasm presence recognition and lesion segmentation, where both predictive quality and inference efficiency are critical for clinical workflows and large-scale screening.

The experimental results show that pruning can substantially reduce the resource footprint of the baseline model, but the achieved speedups and the quality-efficiency trade-off depend strongly on the pruning strategy. In particular, unstructured GMP reduces parameters (from 34.8M to 17.4M at 50% sparsity) but yields only marginal latency improvements (41.5 ms to 40.6 ms), indicating limited hardware benefit without structured sparsity. Structured channel pruning and attention head pruning provide meaningful reductions in FLOPs and latency, yet can introduce a noticeable drop in both classification and segmentation quality at higher sparsity levels.

Across comparable compression regimes, the proposed sensitivity-aware strategy consistently provides a more favorable trade-off. At 50% sparsity, it preserves classification quality better than baselines (F1 = 0.923 vs 0.890 for GMP and 0.906 for structured channel pruning) while maintaining stronger segmentation performance (mIoU = 0.742 vs 0.706 for GMP and 0.734 for structured channel pruning). At the same time, it achieves substantial efficiency gains relative to the unpruned model, reducing latency from 41.5 ms to 23.6 ms and FLOPs from 62.4G to 30.8G. At 60% sparsity, the proposed method retains higher predictive quality than

TABLE III. EFFICIENCY METRICS UNDER PRUNING. LATENCY IS MEASURED ON A FIXED HARDWARE SETUP.

Method	Sparsity	Params (M)	FLOPs (G)	Latency (ms)
Base (Unpruned)	0%	34.8	62.4	41.5
GMP	30%	24.4	62.3	41.0
GMP	50%	17.4	62.1	40.6
Structured Channel	30%	25.1	46.8	32.4
Structured Channel	50%	17.7	34.2	25.3
Structured Channel	60%	13.9	27.1	21.8
Attention Head Pruning	40%	29.2	52.6	35.7
Attention Head Pruning	60%	24.1	44.3	30.1
Proposed Sensitivity-Aware	30%	25.8	44.9	31.2
Proposed Sensitivity-Aware	50%	18.5	30.8	23.6
Proposed Sensitivity-Aware	60%	14.8	23.5	19.4

competing baselines (F1 = 0.904 and mIoU = 0.724) while reaching the best latency among the compared configurations (19.4 ms).

Overall, these results support the conclusion that architecture-aware, sensitivity-guided pruning is an effective direction for compressing attention-augmented lung CT models, enabling large reductions in inference cost while preserving clinically meaningful segmentation and presence recognition performance. Future work will validate the approach on additional CT cohorts and acquisition settings and will explore integration with complementary compression techniques and deployment-oriented optimization objectives.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[2] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *CoRR*, vol. abs/1905.11946, 2019. [Online]. Available: http://arxiv.org/abs/1905.11946

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[5] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," *CoRR*, vol. abs/1711.07971, 2017. [Online]. Available: http://arxiv.org/abs/1711.07971

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015. [Online]. Available: https://arxiv.org/abs/1505.04597

[7] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841517301135

[8] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017. [Online]. Available: https://doi.org/10.1038/nature21056

[9] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, Jan. 2019. [Online]. Available: https://doi.org/10.1038/s41591-018-0300-7

[10] A. Samarin, A. Savelev, A. Toropov, A. Nazarenko, A. Golovatiuk, P. Dmitriev, A. Dzestelova, E. Mikhailova, A. Motyko, and V. Malykh, "Segmentation of the iris and pupil of the human eye in images from an infrared camera," *Pattern Recognition and Image Analysis*, vol. 34, no. 3, pp. 855–862, Sep. 2024. [Online]. Available: https://doi.org/10.1134/S1054661824700743

[11] A. Samarin, A. Toropov, and O. Egorova, "Self-attention based approach to iris segmentation," in *2025 International Russian Smart Industry Conference (SmartIndustryCon)*, 2025, pp. 200–205.

[12] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," 2016. [Online]. Available: https://arxiv.org/abs/1510.00149

[13] T. Gale, E. Elsen, and S. Hooker, "The state of sparsity in deep neural networks," *CoRR*, vol. abs/1902.09574, 2019. [Online]. Available: http://arxiv.org/abs/1902.09574

[14] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," *CoRR*, vol. abs/1608.08710, 2016. [Online]. Available: http://arxiv.org/abs/1608.08710

[15] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, "Learning efficient convolutional networks through network slimming," *CoRR*, vol. abs/1708.06519, 2017. [Online]. Available: http://arxiv.org/abs/1708.06519

[16] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient transfer learning," *CoRR*, vol. abs/1611.06440, 2016. [Online]. Available: http://arxiv.org/abs/1611.06440

[17] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Training pruned neural networks," *CoRR*, vol. abs/1803.03635, 2018. [Online]. Available: http://arxiv.org/abs/1803.03635

[18] V. Sanh, T. Wolf, and A. M. Rush, "Movement pruning: Adaptive sparsity by fine-tuning," *CoRR*, vol. abs/2005.07683, 2020. [Online]. Available: https://arxiv.org/abs/2005.07683

[19] A. Samarin, A. Toropov, A. Dzestelova, A. Nazarenko, E. Kotenko, E. Mikhailova, A. Savelev, and A. Motyko, "Specialized non-local blocks for recognizing tumors on computed tomography snapshots of human lungs," in *2024 35th Conference of Open Innovations Association (FRUCT)*, 2024, pp. 659–664.

[20] A. Samarin, A. Nazarenko, A. Savelev, A. Toropov, A. Dzestelova, E. Mikhailova, A. Motyko, and V. Malykh, "A model based on universal filters for image color correction," *Pattern Recognition and Image Analysis*, vol. 34, no. 3, pp. 844–854, 2024. [Online]. Available: https://doi.org/10.1134/S1054661824700731

[21] A. Samarin, A. Nazarenko, A. Toropov, E. Kotenko, A. Dzestelova, E. Mikhailova, V. Malykh, A. Savelev, and A. Motyko, "Universal filter-based lightweight image enhancement model with unpaired learning mode," in *2024 36th Conference of Open Innovations Association (FRUCT)*, 2024, pp. 711–720.

[22] Y. LeCun, J. Denker, and S. Solla, "Optimal brain damage," in *Advances in Neural Information Processing Systems*, D. Touretzky, Ed., vol. 2. Morgan-Kaufmann, 1989, pp. 598–605. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/1989/file/6c9882bbac1c7093bd25041881277658-Paper.pdf

[23] B. Hassibi, D. Stork, and G. Wolff, "Optimal brain surgeon and general network pruning," in *IEEE International Conference on Neural Networks*, 1993, pp. 293–299 vol.1.

[24] M. Zhu and S. Gupta, "To prune, or not to prune: exploring the efficacy of pruning for model compression," 2017. [Online]. Available: https://arxiv.org/abs/1710.01878

[25] Y. Guo, A. Yao, and Y. Chen, "Dynamic network surgery for efficient dnns," *CoRR*, vol. abs/1608.04493, 2016. [Online]. Available: http://arxiv.org/abs/1608.04493

[26] Y. He, X. Zhang, and J. Sun, "Channel pruning for accelerating very deep neural networks," *CoRR*, vol. abs/1707.06168, 2017. [Online]. Available: http://arxiv.org/abs/1707.06168

[27] J. Luo, J. Wu, and W. Lin, "Thinet: A filter level pruning method for deep neural network compression," *CoRR*, vol. abs/1707.06342, 2017. [Online]. Available: http://arxiv.org/abs/1707.06342

[28] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li, "Learning structured sparsity in deep neural networks," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/file/41bfd20a38bb1b0bec75acf0845530a7-Paper.pdf

[29] V. Lebedev, Y. Ganin, M. Rakhuba, I. Oseledets, and V. Lempitsky, "Speeding-up convolutional neural networks using fine-tuned cp-decomposition," 2015. [Online]. Available: https://arxiv.org/abs/1412.6553

[30] Z. Liu, M. Sun, T. Zhou, G. Huang, and T. Darrell, "Rethinking the value of network pruning," 2019. [Online]. Available: https://arxiv.org/abs/1810.05270

[31] N. Lee, T. Ajanthan, and P. Torr, "SNIP: SINGLE-SHOT NETWORK PRUNING BASED ON CONNECTION SENSITIVITY," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=B1VZqjAcYX

[32] H. Tanaka, D. Kunin, D. L. Yamins, and S. Ganguli, "Pruning neural networks without any data by iteratively conserving synaptic flow," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 6377–6389. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/46a4378f835dc8040c8057beb6a2da52-Paper.pdf

[33] U. Evci, T. Gale, J. Menick, P. S. Castro, and E. Elsen, "Rigging the lottery: Making all tickets winners," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 2943–2952. [Online]. Available: https://proceedings.mlr.press/v119/evci20a.html

[34] T. Yang, A. G. Howard, B. Chen, X. Zhang, A. Go, V. Sze, and H. Adam, "Netadapt: Platform-aware neural network adaptation for mobile applications," *CoRR*, vol. abs/1804.03230, 2018. [Online]. Available: http://arxiv.org/abs/1804.03230

[35] Y. He and S. Han, "ADC: automated deep compression and acceleration with reinforcement learning," *CoRR*, vol. abs/1802.03494, 2018. [Online]. Available: http://arxiv.org/abs/1802.03494

[36] A. Gordon, E. Eban, O. Nachum, B. Chen, T. Yang, and E. Choi, "Morphnet: Fast & simple resource-constrained structure learning of deep networks," *CoRR*, vol. abs/1711.06798, 2017. [Online]. Available: http://arxiv.org/abs/1711.06798

[37] P. Michel, O. Levy, and G. Neubig, "Are sixteen heads really better than one?" in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/2c601ad9d2ff9bc8b282670cdd54f69f-Paper.pdf

[38] The Cancer Imaging Archive (TCIA), "Lidc-idri: Data from the lung image database consortium (lidc) and image database resource initiative (idri)," https://www.cancerimagingarchive.net/collection/lidc-idri/, 2015, data citation required.

[39] S. G. Armato, G. McLennan, L. Bidaut, and et al., "The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans," *Medical Physics*, vol. 38, no. 2, pp. 915–931, 2011.

[40] h. alyasriy and M. AL-Huseiny, "The iq-oth/nccd lung cancer dataset," https://data.mendeley.com/datasets/bhmdr45bh2/4, 2023.

[41] The Cancer Imaging Archive (TCIA), "Lung-pet-ct-dx: A large-scale ct and pet/ct dataset for lung cancer diagnosis," https://www.cancerimagingarchive.net/collection/lung-pet-ct-dx/, 2020, data citation required.

[42] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The cancer imaging archive (tcia): Maintaining and operating a public information repository," *Journal of Digital Imaging*, 2013.