Evaluating a residual learning framework for 3D computed tomography data

Evan Maestri
Stanford University
Department of Immunology
maestri@stanford.edu

Abstract

This work proposes a hybrid deep learning framework for multi-abnormality classification in 3D computed tomography (CT) scans. I introduce a residual learning approach that combines the strengths of 2D ResNet models and 3D Vision Transformers to efficiently analyze volumetric medical imaging data. The framework treats 2D representations as strong baselines with a 3D residual module that integrates across-slice information, addressing the challenge of capturing both slice-level details and volumetric context. In this study, I applied the hybrid approach to a subset of the CT-RATE dataset with labels for 18 lung abnormalities. Our results demonstrate the proposed hybrid approach improves classification accuracy by capturing 3D spatial relationships while maintaining computational efficiency, potentially offering a practical solution for clinical CT image analysis.

1. Introduction

More than 93 million computed tomography (CT) scans are performed annually in the US which are critical in clinical decision making [1]. These detailed three-dimensional (3D) scans often produce hundreds of high-resolution slices providing a comprehensive view of a patient's condition. CT scans are also a preferred modality for diagnostics due to their widespread availability and speed. However, detecting the presence of any abnormalities within the large number of slices of the 3D CT volume requires significant time-consuming interpretation from clinical experts [2].

Deep learning has started to revolutionize our interpretation of medical images [3]. Various radiological imaging tasks have found success using machine learning, such as disease classification, segmentation, lesion detection, and image reconstruction [4, 5]. Despite advances in state-of-the-art vision language models (VLMs), their applications to 3D imaging tasks—such as those involving high resolution volumetric data—remains underexplored. Furthermore, 3D volumetric images can incur large data storage costs and computational complexity due to their high resolution captures which are necessary for radiology diagnostics. Identifying strategies to build high performance models with a fraction of the compute as we utilize large foundation models in healthcare will be crucial [6].

The overall goal of this project is to develop efficient models for multi-abnormality classification in 3D chest CT scans. The input to our algorithm is CT imaging data followed by 2D ResNets or 3D Vision Transformers which output predictions for 18 lung abnormalities. In this work, I propose a hybrid framework that combines the strengths of 2D and 3D deep learning models in a residual learning paradigm. Our approach aims to efficiently capture both slice-level details and volumetric context, improving the efficiency of multi-abnormality classification in CT scans. Key contributions of this study are: A residual learning framework that integrates 2D ResNet and 3D Vision Transformer architectures for CT analysis and evaluation of this hybrid approach on a multi-abnormality CT dataset.

2. Related Work

Given the complexity of 3D CT volumes, a common approach has been to leverage 2D convolutional neural networks (CNNs) for CT analysis, treating each slice as an independent image. For instance, 2D CNNs have been used to detect different subtypes of intracranial hemorrhages in head CT scans [7]. In addition, models to detect thoracic abnormalities via CNNs leveraging detailed insights into complex anatomical structures from CT scans have also performed well [8]. The success of deep learning in imaging has been driven in large part by ResNet architectures [9]. ResNets have found widespread use due to their ability to train very deep networks effectively through residual connections. While deeper models offer more representational power due to their increased number of parameters, they also present greater optimization challenges. The residual blocks make deeper networks easier to optimize because they mitigate the problem of vanishing gradients by using shortcut connections that perform identity mapping. Rather than learning unreferenced functions, this residual approach

enables the network to effectively skip layers that do not contribute to improving performance. However, while 2D approaches have shown promise, they often struggle to fully capture the spatial relationships present in 3D CT volumes [10].

To address this limitation, researchers have developed 3D architectures specifically designed for volumetric data analysis. Vision Transformers (ViTs) are an architectural adaptation of transformers [11], originally developed for natural language processing, to address core computer vision challenges. Given their rapid success in 2D vision tasks, ViTs have been extended to more complex modalities, such as volumetric medical data [12]. This has lead to the recent development of architectures specifically designed for 3D medical imaging, such as the CT Vision Transformer (CTViT) [13] which encodes 3D CT volumes into tokens. The model was inspired by video transformers (e.g., ViViT [14]) and thus extracts spatiotemporal tokens where instead of the time dimension it uses the depth dimension (or slice dimension) of the CT volume.

Despite the advancements of ViTs, CNNs have retained their relevance due to their computational efficiency, lower complexity, smaller parameter space, and competitive performance on smaller datasets. Combining attention mechanisms with CNN architectures in various forms, has been an active area of work to retain maximal benefits of each model strength [15, 16].

3. Data

The dataset I used is CT-RATE [17] which consists of non-contrast 3D chest CT scans from 21,304 patients totalling over 14.3 million 2D slices. I sampled from this dataset approximately 3000 patients and 2 million slices. The dataset has multi-abnormality labels for 18 lung conditions (see Appendix).

3.1. Pre-processing

CT scans are loaded from NIFTI (.nii.gz) files. To standardize the CT volumes, they are either center-cropped or padded to reach a consistent resolution of 214x214x240. Raw pixel values are converted to Hounsfield units (HU) with clipping to a range of [-1000,1000] HU. During training, the values are normalized to a range of -1 to 1.

A center sampling method was used to extract a subset of slices from the middle of the volume (num_slices=100), with a corresponding volume of 214x214x100 per CT scan. Each volume has corresponding multi-abnormality labels.

For training, the dataset was split 70:20:10 into a training, validation, and test set using stratified sampling to ensure equitable distributions of the classes in each split. The class distributions within each split can be seen in Table 1 with example CT images shown in Figure 1.

Class	Total	Train	Val	Test
Medical material	425	285	97	43
Arterial wall calcification	909	604	198	107
Cardiomegaly	358	240	76	42
Pericardial effusion	269	174	65	30
Coronary arterial calcification	793	541	163	89
Hiatal hernia	425	287	94	44
Lymphadenopathy	831	540	189	102
Emphysema	586	392	127	67
Atelectasis	842	561	192	89
Lung nodule	1393	965	282	146
Lung opacity	1141	782	242	117
Pulmonary fibrotic sequela	783	531	178	74
Pleural effusion	481	316	113	52
Mosaic attenuation pattern	268	176	62	30
Peribronchial thickening	324	214	77	33
Consolidation	569	374	129	66
Bronchiectasis	307	209	68	30
Interlobular septal thickening	273	179	63	31

Table 1: Distribution of classes across data split.

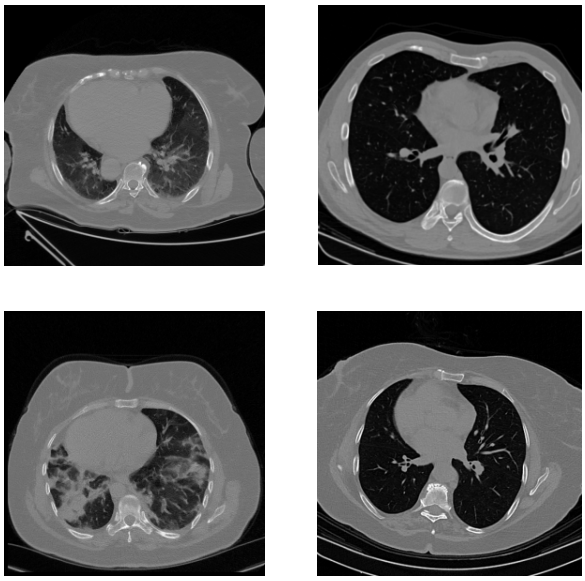

Figure 1: Representative lung CT images with various lung pathologies including lung nodule, atelectasis, lung opacity, and consolidation.

4. Methods

Inspired by the residual learning paradigm introduced in ResNet [9], the general framework proposed here is that models should be able to learn as least as well as a base 2D representation, with residual updates coming from the 3D volumetric context. This involves (1) a hybrid framework that treats 2D representations as strong baselines which

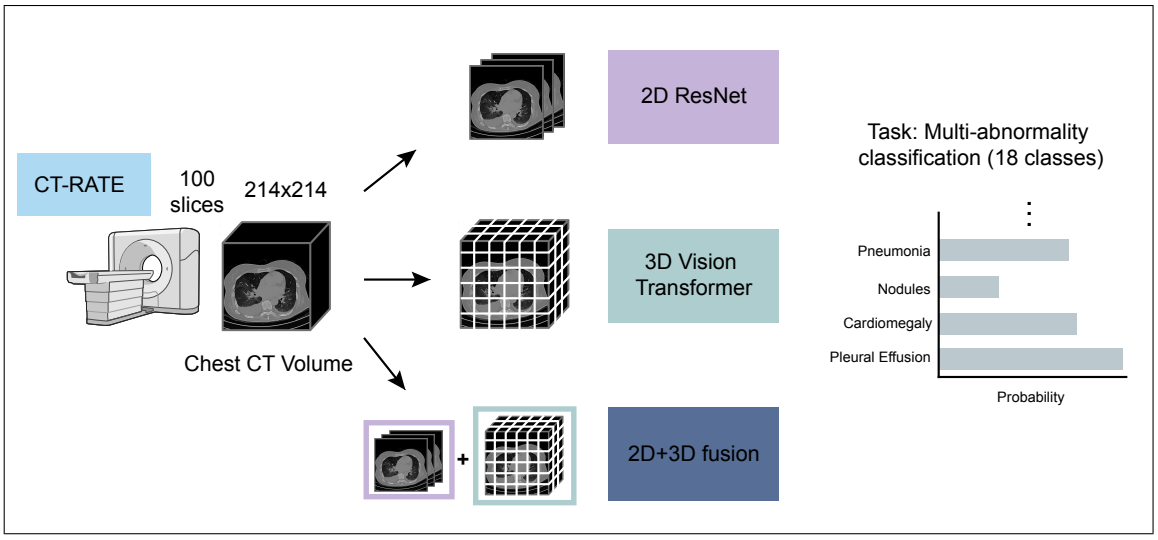

Figure 2: Pipeline overview and computational framework.

will require a base 2D encoder capturing features per slice and (2) a 3D residual module which can integrate across-slice information to refine updates from the added volumetric layer (Figure 2). The implementation will be a 2D ResNet+3D ViT fusion. Training 3D models can be compute-heavy, so by leveraging strong pretrained 2D models with lightweight or modular 3D enhancement we may reduce the computational overhead.

4.1. 2D Models: ResNet

The 2D ResNet model serves as the baseline approach for the CT multi-label abnormality classification task. This 2D approach processes each of the 100 slices (214×214 pixels) from our CT volumes independently, treating them as separate grayscale images. This approach can effectively give a baseline for learning patterns within individual slices. Here, we process the CT data slice-by-slice (a 2D approach), using a modified ResNet-18 architecture [9] pre-trained on ImageNet. The model has been adapted for this task by: 1) Modifying the input layer to accept single-channel grayscale images (CT slices), replacing the standard RGB three-channel input. 2) Replacing the final fully connected layer to output per-slice predictions for 18 distinct classes for the CT multi-label abnormality task. The set of predictions per-slice indicate the likelihood of each abnormality being present in that slice or not. The slice-level predictions are aggregated (mean) across the entire volume to generate a final multi-label prediction per CT scan reflecting if the abnormality is present anywhere in the volume.

4.2. 3D Model: ViT

ViTs work by breaking images into patches (analogous to word tokens in language models), which can then be processed in a series of transformer encoder blocks. Positional encoding are incorpororated to preserve spatial information before passing the embedded patches into the transformer encoder stack. Each encoder block consists of a multi-head self-attention mechanism that enables the network to focus on different image regions simultaneously, followed by a feedforward neural network and a LayerNorm (RMSNorm) operation. The encoder's output is fed into a classification head, a multi-layer perceptron (MLP), which generates class probability scores for the target categories.

We adapt the CT Vision Transformer (CTViT) [13] architecture to encode the 3D CT volumes into tokens. Our implementation of ViT accepts input volumes of size 224x224x100 (representing 100 slices of 224x224 pixels each). The encoder network (ϕ_e^{ViT}) processes this input to produce embedded CT tokens, while the decoder network (ϕ_d^{ViT}) uses these tokens for classification tasks.

The encoder first extracts non-overlapping patches of 16x16 pixels from each slice of the CT volume. These patches are linearly transformed into a 512-dimensional latent space. For a batch of CT volumes, this results in a tensor of shape $B \times T \times (H/p_1) \times (W/p_2) \times D$, where B is the batch size, T is the number of slices (100 in our case), H and W are the height and width of the slices (224 each), p_1 and p_2 are the spatial patch sizes (16 each), and D is the latent dimension (512).

This tensor is then processed by two transformer networks in sequence. The spatial transformer operates on the spatial dimensions, while the slice-wise transformer pro-

cesses information across slices. Both transformers use 8 attention heads with a dimension of 64 per head. The architecture includes 4 spatial depth layers and 4 slice-wise depth layers, allowing it to capture both intra-slice and inter-slice relationships. While the original CTViT was designed for volume reconstruction, our ViT version is adapted for multi-label classification. The decoder network in our implementation is modified to output classification predictions for the 18 abnormality classes. The ViT model was trained from scratch without utilizing any pre-trained weights.

4.3. Residual Fusion 2D+3D Model

The hybrid framework proposed here adapts the idea of residual blocks from ResNet: $H(x) = F(x) + x$ as outlined in Algorithm 1. The fusion process involves the 2D ResNet processing the input CT slices, producing a set of slice-level feature vectors. Then, the slice-level features are aggregated via mean pooling to obtain a single feature vector per volume. Concurrently, the 3D ViT processes the same input CT volume, encoding the volumetric context into another set of feature vectors. A projection layer ensures dimensionality matching between the two feature spaces, mapping the 3D ViT features to the same dimensionality as the 2D ResNet features. The outputs of the 2D ResNet and 3D ViT base models are then combined through element-wise addition. The final fused representation, is then passed through a classification head to produce predictions for the 18 abnormality classes. This residual connection allows the ViT to learn to correct or enhance the 2D ResNet's predictions, rather than being forced to learn everything from scratch.

Algorithm 1: Residual fusion from 2D & 3D inputs

Input: $input_{2d}, input_{3d}$
Output: $output_{fusion}$
$x_{2dResNet} \leftarrow model_{2dResNet}(input_{2d})$;
$residual_{3dViT} \leftarrow model_{3dViT}(input_{3d})$;
$output_{fusion} \leftarrow x_{2dResNet} + residual_{3dViT}$;

4.4. Evaluation method

The overall task is multi-abnormality classification (18 classes). Accuracy, precision, recall, and AUROC scores were used for evaluation metrics on the test data as implemented in sklearn.metrics.

4.5. Experimental details

For this multi-label classification task, the loss function used was binary cross entropy loss (BCEWithLogitsLoss), as implemented in PyTorch. To easily compare the performance of multiple models they were built using a PyTorch training loop for experiments with the following parameters kept the same to fine-tune the models and make direct comparisons: learning_rate=5e-4, num_train_epochs=10, optimizer=AdamW, batch_size=8, lr_scheduler_type = ReduceLROnPlateau, which monitors the validation loss and reduces the learning rate by a factor of 0.5 if no improvement is seen after 2 consecutive epochs to help prevent overfitting. The training time was $\approx$ 8 hours per model.

4.6. Model training environment

Training was performed on a multi-GPU system equipped with 4 NVIDIA H100-80GB-HBM3 GPUs. Distributed data parallelism was implemented using PyTorch's DistributedDataParallel (DDP) module, with data split along the batch dimension across the 4 GPUs to accelerate training.

5. Results and Discussion

The performance of the baseline 2D method ResNet18, consisting of 18 layers demonstrated a reasonable starting discrimination capability with an AUC of 0.83 and an accuracy of 0.83 (Table 2). The model achieved a precision of 0.66, meaning when it predicts a positive it is correct about two-thirds of the time and a low recall at 0.34 indicating it is missing two-thirds of the actual positives. This means the model is good at distinguishing classes, when it predicts positives it is often right (high precision); however, it is conservative in predicting positives (low recall).

Next, the baseline 3D method ViT achieved a lower performance compared with the ResNet with an AUC of 0.71, accuracy of 0.75, precision of 0.44 and recall of 25. This performance gap is likely due to two factors: the ResNet benefits from ImageNet pre-training which provides useful feature representations for our CT images, while our ViT was trained from scratch. Additionally, the 3D ViT model has more parameters to optimize with our relatively small training dataset size.

Finally, the fusion model combining both 2D and 3D models (ResNet+ViT) showed substantial improvement over both individual models (Figure 3), with AUC increasing to 0.87, accuracy to 0.87, precision to 0.72 and recall to 0.54. This significant enhancement in recall (0.34 ResNet to 0.54 ResNet+ViT) while maintaining high precision indicates that the 3D volumetric context is helping identify abnormalities that would be missed when examining the slices in isolation.

Model	2D	3D	AUC	Accuracy	Precision	Recall
ResNet	✓		0.83	0.83	0.66	0.34
ViT		✓	0.71	0.75	0.44	0.25
ResNet+ViT	✓	✓	0.87	0.87	0.72	0.54

Table 2: Comparison of model performance.

Next, I performed experiments on the fusion model. In the initial implementation in Table 2, equal weights were

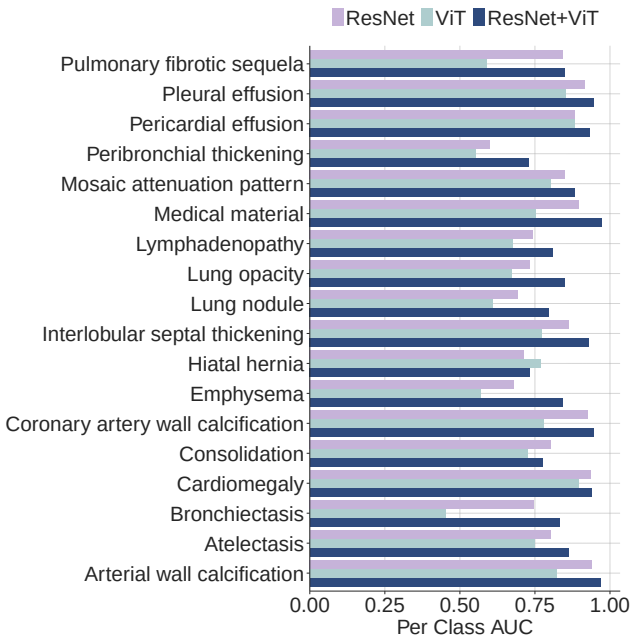

Figure 3: Per class AUC on unseen test data.

given to both the ResNet and the ViT. Here, I added an α parameter to control the fusion from the 2D and 3D sources. The update to Algorithm 1 is as follows:

$$output_{fusion} \leftarrow x_{2dResNet} + \alpha \cdot residual_{3dViT}$$

Using an $\alpha = 0.0$ corresponds to using only ResNet features, while $\alpha = 1.0$ gives equal weight to both ResNet and ViT features. Using the learned α approach allows the model to adaptively determine the optimal contribution from the 3D ViT model component. By making the α a learnable parameter, it allows the model to balance the 2D/3D fusion representation. Thus, the model could effectively learn to set the alpha close to 0 during training (giving negligible weight to the 3D residual) if the 3D features don't contribute positively to the classification task. On the other hand, if the 3D ViT features contain complementary information to the ResNet, the alpha would be optimized to a higher value, appropriately weighting the contribution of both modalities.

We can observe from Table 3 that as we increase the alpha parameter, the overall performance increases across all metrics. In particular, we see strong increases in the recall (from 0.34 to 0.54), which means the model is better able to detect true positive cases. Thus, this indicates that the combined feature representation including both 2D ResNet and 3D ViT features is providing valuable complementary information. Moreover, the experiment with the learned alpha indicates that the model is able to effectively tune contributions from both modalities, achieving similar performance

α	AUC	Accuracy	Precision	Recall
0.0	0.83	0.83	0.66	0.34
0.1	0.85	0.85	0.7	0.36
0.3	0.86	0.86	0.72	0.46
0.5	0.88	0.87	0.75	0.54
1.0	0.87	0.87	0.72	0.54
learned	0.87	0.87	0.75	0.55

Table 3: Performance of the fusion models (ResNet+ViT) with varying α parameter values.

to the experiments with the $\alpha = 0.5, 1$.

Given the flexibility that comes from using a learnable model fusion weighting, I ran a final fusion model increasing to the full CT image height/width size of 480x480. On this larger model, I examined the latent embedding space with t-SNE (t-Distributed Stochastic Neighbor Embedding), coloring by the number of abnormalities present in the CT volume (Figure 4). We can observe a clustering of abnormalities, suggesting that while the model was utilized for multi-abnormality detection, it could also help in prognosis/disease grading. For example, future research could be used to project new CT scans into the latent space of a larger trained model for better clustering of lung disease endotypes. This approach could also support retrieval applications, enabling clinicians to input a new patient scan and identify similar cases in a database to help inform treatment decisions by referencing outcomes from patients with comparable disease patterns.

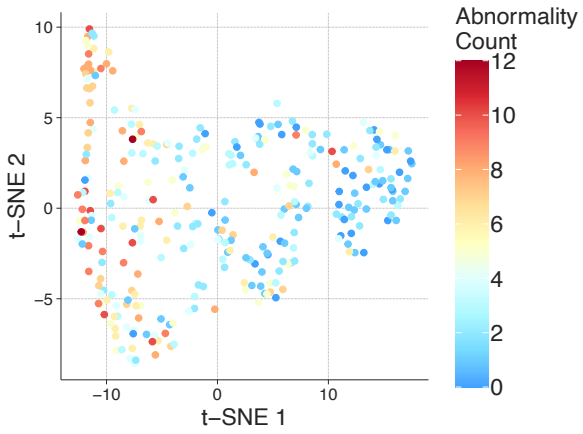

Figure 4: The t-SNE latent space colored by the number of abnormalities present in the CT volume.

I also visualized the cross-attention maps to better understand model interpretability examining how these models were making classification decisions. Figure 5 demon-

strates that the attention maps from the fusion model is picking up the underlying lung morphology structure. However, the model seems to struggle with pathology where the visualization shows it is more diffuse throughout the lung, such as mosaic attenuation pattern (top right of Figure 5. In contrast, the lung nodule (a more localized pathology, bottom right of Figure 5) can be seen to have two regions in the image which have higher attention weights. This suggests that the model's attention mechanism is more effective at identifying localized abnormalities with distinct boundaries compared to diffuse patterns that affect larger portions of the lung tissue. The ViT model's stronger visual representation on nodules likely stems from the training class imbalance, which provided the model with significantly more examples of lung nodules ($\approx$ 4x) to learn the distinctive features of compared with smaller class categories. Since the mosaic attenuation pattern did not have worse performance at a per class AUC level, this suggests that the ResNet features may contribute more to performance for this class label.

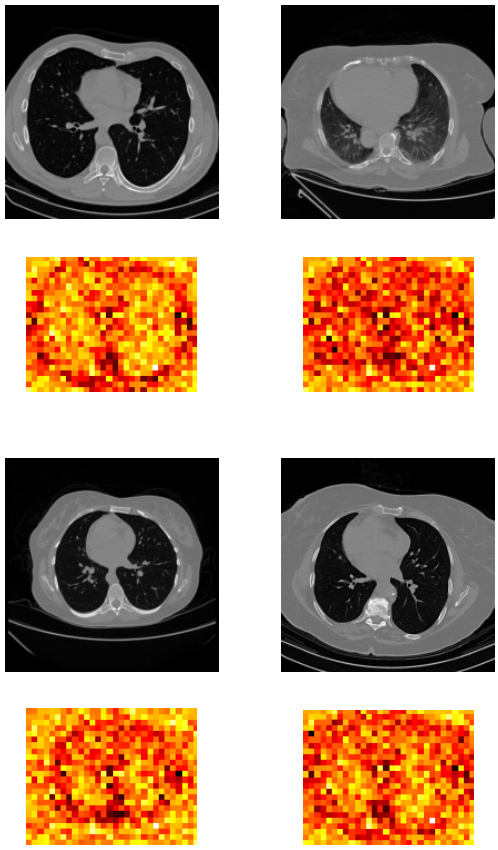

Figure 5: Cross-attention maps for showing specific abnormalities, highlighting the ability of the model to determine relevant regions.

6. Conclusion/Future Work

In this work, I presented a residual learning framework that successfully combines 2D ResNet and 3D Vision Transformer architectures to improve multi-abnormality classification in 3D chest CT scans. The algorithm which was the highest-performing was the 2DResNet+3DViT fusion model. Incorporating the 3D volumetric context helps identify abnormalities that would be missed when examining slices in isolation. This confirms our hypothesis that the residual learning framework allows the 3D component to effectively augment the strong 2D representation rather than learn everything from scratch.

The underperformance of the standalone 3D ViT model likely stems from our relatively small dataset size and lack of pre-training, highlighting the advantage of our hybrid approach in leveraging pre-trained 2D models while still incorporating 3D context. Thus, I expect future iterations with larger training sizes using the full CT-RATE dataset will improve the ViT performance to learn a better representation.

Overall, I examined different architectural approaches for leveraging the strengths of combining both 2D and 3D models for imaging analysis. The CT-RATE dataset also has paired radiology reports (text data), which was not analyzed in the current study. Thus, in future work I would anticipate adding in the additional text dimension which should further increase model performance. Lastly, I would anticipate that a learned weighting parameter for the 2D ResNet features could be also added such that instead of only varying the 3D update, the 2D feature contribution could also be adaptively controlled.

7. Appendix

The CT-RATE dataset is publicly available for download here: https://huggingface.co/datasets/ibrahimhamamci/CT-RATE. Code which was used to adapt our ViT implementation can be found here: https://github.com/ibrahimethemhamamci/CT2Rep.

The CT-RATE dataset has multi-abnormality labels for the following 18 classes: Medical material, Arterial wall calcification, Cardiomegaly, Pericardial effusion, Coronary artery wall calcification, Hietal Hernia, Lymphadenopathy, Emphysema, Atelectasis, Lung nodule, Lung opacity, Pulmonary fibrotic sequela, Pleural effusion, Mosaic attenuation pattern, Peribronchial thickening, Consolidation, Bronchiectasis, Interlobular septal thickening.

8. Contribution and Acknowledgements

Evan Maestri designed experiments, trained models, evaluated performance, and wrote the report. Course instructor Matthew Jin advised on technical details and implementation.

References

[1] R. Smith-Bindman, P. W. Chu, H. Azman Firdaus, C. Stewart, M. Malekhedayat, S. Alber, W. E. Bolch, M. Mahendra, A. Berrington de González, and D. L. Miglioretti, "Projected lifetime cancer risks from current computed tomography imaging," *JAMA Intern. Med.*, Apr. 2025.

[2] A. Udare, M. Agarwal, K. Dhindsa, A. Alaref, M. Patlas, A. Alabousi, Y. K. Kagoma, and C. B. van der Pol, "Radiologist productivity analytics: Factors impacting abdominal pelvic CT exam reporting times," *J. Digit. Imaging*, vol. 35, pp. 87–97, Apr. 2022.

[3] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher, "Deep learning-enabled medical computer vision," *NPJ Digit. Med.*, vol. 4, p. 5, Jan. 2021.

[4] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," 2019.

[5] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D. P. Naidich, and S. Shetty, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nat. Med.*, vol. 25, pp. 954–961, June 2019.

[6] M. Varma, A. Kumar, R. van der Sluijs, S. Ostmeier, L. Blankemeier, P. Chambon, C. Bluethgen, J. Prince, C. Langlotz, and A. Chaudhari, "Medvae: Efficient automated interpretation of medical images with large-scale generalizable autoencoders," 2025.

[7] X. Wang, T. Shen, S. Yang, J. Lan, Y. Xu, M. Wang, J. Zhang, and X. Han, "A deep learning algorithm for automatic detection and classification of acute intracranial hemorrhages in head CT scans," *NeuroImage Clin.*, vol. 32, p. 102785, Aug. 2021.

[8] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1207–1216, 2016.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[10] Q. Dou, L. Yu, H. Chen, Y. Jin, X. Yang, J. Qin, and P.-A. Heng, "3D deeply supervised network for automated segmentation of volumetric medical images," *Med. Image Anal.*, vol. 41, pp. 40–54, Oct. 2017.

[11] 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," 2021.

[12] L. Blankemeier, J. P. Cohen, A. Kumar, D. V. Veen, S. J. S. Gardezi, M. Paschali, Z. Chen, J.-B. Delbrouck, E. Reis, C. Truyts, C. Bluethgen, M. E. K. Jensen, S. Ostmeier, M. Varma, J. M. J. Valanarasu, Z. Fang, Z. Huo, Z. Nabulsi, D. Ardila, W.-H. Weng, E. A. Junior, N. Ahuja, J. Fries, N. H. Shah, A. Johnston, R. D. Boutin, A. Wentland, C. P. Langlotz, J. Hom, S. Gatidis, and A. S. Chaudhari, "Merlin: A vision language foundation model for 3d computed tomography," arXiv 2024.

[13] I. E. Hamamci, S. Er, A. Sekuboyina, E. Simsar, A. Tezcan, A. G. Simsek, S. N. Esirgun, F. Almas, I. Dogan, M. F. Dasdelen, C. Prabhakar, H. Reynaud, S. Pati, C. Bluethgen, M. K. Ozdemir, and B. Menze, "Generatect: Text-conditional generation of 3d chest ct volumes," 2024.

[14] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," 2021.

[15] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *CoRR*, vol. abs/2005.12872, 2020.

[16] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. J. Smola, "Resnest: Split-attention networks," *CoRR*, vol. abs/2004.08955, 2020.

[17] I. E. Hamamci, S. Er, F. Almas, A. G. Simsek, S. N. Esirgun, I. Dogan, M. F. Dasdelen, O. F. Durugol, B. Wittmann, T. Amiranashvili, E. Simsar, M. Simsar, E. B. Erdemir, A. Alanbay, A. Sekuboyina, B. Lafci, C. Bluethgen, M. K. Ozdemir, and B. Menze, "Developing generalist foundation models from a multimodal dataset for 3d computed tomography," 2024.