AIMI2024 Project: LUNA23

Antonio Carpes
Radboud University
s1132642
antonio.carpes@ru.nl

Andreas Pattichis
Radboud University
s1132644
andreas.pattichis@ru.nl

Honor Duthie
Radboud Unviersity
s1132643
honor.duthie@ru.nl

Stanislav Gergert
Radboud University
s1050772
stanislav.gergert@ru.nl

***Abstract*—This project aimed to improve the LUNA23 baseline results for lung nodule analysis using advanced models. We implemented nnU-Net for segmentation and a Vision Transformer (ViT) for nodule type and malignancy classification. Despite using data balancing techniques, the models faced overfitting and did not exceed baseline performance. nnU-Net achieved a validation dice score of 0.823 but underperformed on the test set, while ViT showed low balanced accuracy of 0.622 in validation and was not applied to the test set.**

***Index Terms*—Lung cancer, 3D-CNN, U-Net, nnU-Net, Vision Transformer, Segmentation, Classification**

I. Introduction

Our project, LUNA23, works with isolated computed tomography (CT) scans of isolated lung nodule to: classify nodule malignancy, classify nodule type, and apply nodule segmentation, in order to help identify lung cancer. Our goal was to experiment with different machine learning techniques to attempt to improve upon the baseline model results.

II. Methods and Materials

1) ***3D Convolutional Neural Network (3D-CNN):*** The 3D-CNN was used for both malignancy and nodule type classification, pretrained on the Kaggle Data Science Bowl 2017 dataset [1]. It processes spatial hierarchies within CT scans using multiple convolutional layers with ReLU activation and max-pooling. The final dense layer outputs classifications using Sigmoid for malignancy and Softmax for nodule type. The 3D-CNN captures 3D spatial information crucial for distinguishing texture and internal opacity in lung nodules, but it requires significant hardware resources and extensive labelled data [1].

2) ***U-Net:*** U-Net, introduced by Ronneberger et al. (2015), was used for nodule segmentation [2]. It consists of a contracting path to capture context and an expansive path for precise localisation, using 3D convolutions, ReLU activations, batch normalisation, and max-pooling. The final output layer uses a Sigmoid activation function to generate segmentation masks. U-Net efficiently captures fine details and spatial context but requires substantial computational resources and manual hyperparameter tuning [2].

3) ***nnU-Net:*** To improve U-Net's limitations, we implemented nnU-Net for segmentation [3]. nnU-Net automates the configuration process, adapting preprocessing, architecture selection, and training schedules to the dataset, reducing the need for manual tuning and enhancing segmentation accuracy. It employs a multi-stage architecture with 3D convolutions, instance normalisation, and leaky ReLU activation. nnU-Net requires significant computational resources, and its adaptability may increase initial setup complexity [3].

4) ***Vision Transformer (ViT):*** To enhance nodule type classification, we implemented the Vision Transformer model *vit_base_patch16_224* [4]. Unlike traditional CNNs, ViT uses a transformer architecture, processing images by splitting them into fixed-size patches and using a transformer encoder with self-attention mechanisms. ViTs capture global contextual information effectively and leverage pre-training on large datasets, improving performance with smaller medical datasets. However, they are computationally intensive and require substantial memory and training data [4].

A. Dataset Description and Characteristics

The dataset originates from the Lung Nodule Analysis 2023 - ISMI educational challenge, evaluating algorithms for lung nodule analysis using chest CT images. The challenge includes three tasks: malignancy risk estimation, nodule type classification, and nodule segmentation.

The dataset is divided into training and testing sets. The training set consists of 687 lung nodule images with pixel-level labels for segmentation and a CSV file with labels for nodule type and malignancy. There were no missing values in the CSV. The testing set includes 256 lung nodule images without pixel-level labels or CSV files for nodule type and malignancy. Participants generate these labels, with a hidden testing dataset used for final evaluation through automatic assessment.

Key attributes of the dataset include:

- Each nodule's volume of interest (VOI) is 128 x 128 x 64 voxels.
- Nodule types are classified into non-solid, part-solid, solid, and calcified, labeled as classes 0, 1, 2, and 3.
- Malignancy risk is indicated by binary labels: 0 for non-malignant and 1 for malignant nodules.
- Binary voxel-level labels are provided, with class 0 for the background and class 1 for the nodule.

TABLE I: Distributions of nodule types and malignancy labels, indicating some class imbalances

Nodule Type (%)	**Malignancy (%)**
Solid (67.83)	Non-malignant (69.14)
Calcified (22.13)	Malignant (30.86)
GroundGlassOpacity (7.42)	-
SemiSolid (2.62)	-

Pixel intensity comparison between training and test datasets is summarised in Table II. Analysis shows similar pixel intensities in the training and test datasets, with slight variations in mean and standard deviation. This consistency suggests that the datasets are comparable, which benefits model training and evaluation.

TABLE II: Pixel intensities for training and test images.

Statistic	Training Images	Test Images
Mean	-442.25 ± 150.51	-447.01 ± 142.18
Standard Deviation	462.47 ± 89.07	465.38 ± 86.10
Minimum Intensity	-1235.90	-1272.59
Maximum Intensity	1679.21	1744.77

B. Experimental Approaches

*1) **Baseline Approach**:* In the baseline approach, we trained 3D-CNN for malignancy and nodule type classification and U-Net for segmentation.

*2) **Second Approach**:* To improve segmentation results, we trained nnU-Net for segmentation while keeping 3D-CNN for classification tasks.

*3) **Third Approach**:* We implemented Vision Transformer to improve nodule type classification and used the baseline models for other tasks to leverage the global context capturing ability of ViTs.

*4) **Fourth Approach**:* To address the significant class imbalance more effectively, we complemented the use of the Weighted Random Sampler with additional data preprocessing steps. Specifically, we created balanced training and validation sets for all 5 folds, ensuring both sets within each fold were balanced and independent. We applied various data augmentation techniques to the minority classes of each set until their proportions matched the majority class. These techniques included:

- **Rotation:** Random angles between -30 and 30 degrees.
- **Translation:** Shifts in the x, y, and z directions within a range of -5 to 5 pixels.
- **Flipping:** Horizontal, vertical, and depth-axis flips.
- **Scaling:** Random factors between 0.9 and 1.1.

Each approach was assessed by training and validation metrics. The best-performing models from each approach were then tested and compared using inference to determine the best model for each task.

C. Training and Validation

During the training process of each model, we employed *Stratified K-Fold* for balanced splitting of the dataset into training and validation sets. This ensured that each fold in cross-validation had the same class distribution as the overall dataset. To address class imbalance during training, we used *Weighted Random Sampler*, which adjusts sampling probabilities based on class frequencies, ensuring balanced representation of all classes in each training batch. The sampler was integrated into the data loader for consistent balance.

We monitored training and validation loss, as well as the following metrics: **balanced accuracy** for nodule type classification, **Dice coefficient** for segmentation, **Area under the ROC curve (AUC)** for malignancy classification, and **Pseudo Dice score** for nnU-Net segmentation.

TABLE III: Training Settings for Different Models

Parameter	Settings
Epochs	Baseline: 1000 (3D-CNNs), 100 (U-Net), Second: 1000 (nnU-Net), Third: 1000 (ViT), Fourth: ≈ 200 (Nodule Type), ≈ 80 (Malignancy)*
Batch size	32 (classification), 4 (segmentation), 5 (nnU-Net)
Optimizer	Adam (α: 1×10^{-4}, 0.01 for nnU-Net)
Loss functions	Binary cross-entropy (malignancy), Categorical cross-entropy (nodule type), Dice loss (segmentation)

[a]Trained on with fewer epochs, due to Snellius' long queue times.

During inference, we loaded the trained models and performed predictions on the test set. The pre-processing involved loading and scaling the images, followed by model predictions for segmentation, malignancy, and nodule type classification. We then post-processed segmentation outputs by resampling to original spacing, padding, cropping, and thresholding to obtain binary masks. To ensure accuracy, we retained only the central connected component of the segmented nodules. The performance of the models during inference was evaluated using the same metrics as during training and validation to ensure consistency.

III. Results

A. Evaluation Methods

We evaluated each approach using training and validation loss and score plots, inference scores on the test set, summary statistics for validation predictions, and Grad-CAM (Gradient-weighted Class Activation Mapping) visualisations. For each task, we plotted training and validation losses to monitor learning curves, identified the best validation scores to select optimal models, and performed inference to obtain final prediction scores for comparison. Summary statistics were compiled for comprehensive evaluation. Grad-CAM visualisations, using the last convolutional layer for 3D-CNN, the last layer in the contraction path for U-Net, and the last attention block for Vision Transformer (ViT), were generated to identify regions in the CT images that contributed most to the predictions, providing insights into the model's decision-making process.

B. Key Findings

1) Training and Validation Losses and Scores: The training and validation losses and scores for each model were plotted to evaluate their convergence and performance.

In the **Baseline Approach**, the U-Net model's training and validation losses indicate successful learning and convergence without substantial overfitting, suggesting strong model performance. In contrast, the 3D-CNN model for malignancy classification shows significant fluctuations in validation loss, suggesting instability (Figure 1). However, the point of model selection, marked by a red dot, is reached before these fluctuations intensify, enabling the model to achieve a high AUC.

In the **Second Approach**, the nnU-Net achieved a mean validation Dice score comparable to the U-Net, highlighting its robustness in segmentation tasks.

In the **Third Approach**, the Vision Transformer for nodule type classification exhibits consistent training performance but lower validation accuracy, indicating challenges in generalizing from the training data.

In the **Fourth Approach**, models trained on balanced datasets demonstrated improved stability in their learning curves (example in Figure 2).

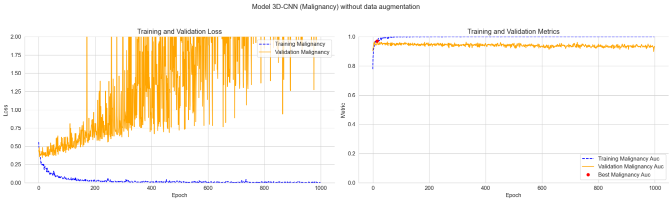

Fig. 1: 3D-CNN (Malignancy) without data augmentation.

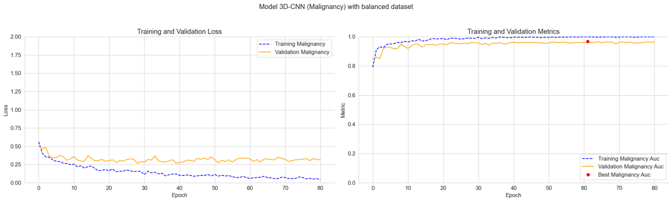

Fig. 2: 3D-CNN (Malignancy) with balanced dataset.

The table IV summarises the best metrics for each model, with the highest validation performance scores for each task highlighted in green:

TABLE IV: Best Validation Performance Metrics per Model

Model	Best Metric
U-Net (Segmentation)	Dice: 0.823
nnU-Net (Segmentation)	Mean Validation Dice: 0.823
3D-CNN (Nodule Type, No Balancing)	Balanced Accuracy: 0.821
3D-CNN (Nodule Type, Balancing)	Balanced Accuracy: 0.855
Vision Transformer (Nodule Type)	Balanced Accuracy: 0.622
3D-CNN (Malignancy, No Balancing)	AUC: 0.972
3D-CNN (Malignancy, Balancing)	AUC: 0.967

2) Grad-CAM Analysis: The Grad-CAM visualisations provided insights into the regions of the CT images that influenced the models' predictions. These visualisations were crucial in validating the models' focus on relevant anatomical structures.

In the **Third Approach**, the Vision Transformer for nodule type classification exhibits a focus that is more distributed throughout the image compared to the 3D-CNN in the baseline approach, due to the global attention mechanism of ViT (Figure 3).

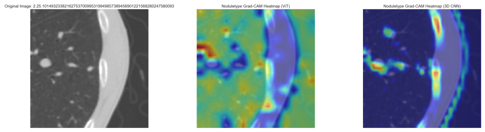
Original Image: 2.25.10149323382162753700995319949857389456901221568280247580093
Noduletype Grad-CAM Heatmap (ViT)
Noduletype Grad-CAM Heatmap (3D CNN)

Fig. 3: **Approach 3**: Grad-CAM heat maps for Vision Transformer (ViT) and 3D-CNN in nodule type classification.

In the **Fourth Approach**, the models trained on balanced datasets demonstrate a wider focus in their Grad-CAM heat maps compared to the unbalanced datasets, as shown in the sample for nodule type classification (Figure 4).

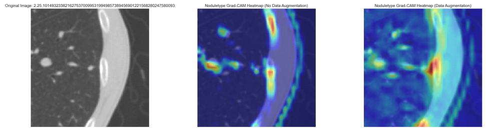
Original Image: 2.25.10149323382162753700995319949857389456901221568280247580093
Noduletype Grad-CAM Heatmap (No Data Augmentation)
Noduletype Grad-CAM Heatmap (Data Augmentation)

Fig. 4: **Approach 4**: Grad-CAM heat maps for nodule type classification without and with data augmentation.

3) Performance on Test Set: Inference was conducted to evaluate the final model performance on the test set, using the same metrics as during training and validation to ensure consistency. Various model combinations were tested to enhance performance, but none surpassed the baseline. **Despite some models performing well on the validation set, their test set performance did not show significant improvements**. **ViT model was excluded from the inference stage** due to poor validation results, indicating unsuitability for our dataset and tasks. Some of the model combinations tested for their performance are as follows:

1) Baseline model
2) 3D-CNN model for malignancy prediction (trained on a balanced dataset, with other models the same as baseline)
3) nnU-Net model for segmentation (with other models the same as baseline)
4) Combination of nnU-Net for segmentation, 3D-CNN for noduletype prediction on a balanced dataset, and 3D-CNN for malignancy prediction (baseline)

TABLE V: Test performance metrics of different model combinations.

Model	Overall Metric	Segmentation Dice	Malignancy Risk AUC	Nodule Type Accuracy
1	0.79	0.646	0.847	0.82
2	0.795	0.652	**0.862**	0.805
3	0.743	**0.434**	0.867	0.805
4	0.668	0.434	0.867	**0.508**

These results suggest that while the **3D-CNN model for malignancy prediction** showed some promise, overall improvements were not substantial. The **nnU-Net model for segmentation** under performed, leading to a drop in segmentation performance. Further optimisation and exploration of model combinations are needed to achieve significant improvements over the baseline.

C. Final Model Selection

As the purpose of this study was not just to get the best score but also to test novel combinations, we selected a combination of models that showed some good results among the other combinations tested. This combination, while not better than the baseline, provided valuable insights. The selected combination included nnU-Net for segmentation, 3D-CNN for malignancy trained on a balanced dataset, and the baseline model for nodule type. The outcomes align with expectations: there is no enhancement or deterioration in the malignancy balanced model, the nodule type model remains stable, and the performance of nnU-Net is inferior to that of the original U-Net:

- **Nodule Type Classification:** Baseline Model
- **Segmentation:** nnU-Net
- **Malignancy Classification:** 3D-CNN (trained on balanced dataset)

IV. Discussion

The purpose of this project was to attempt to improve the LUNA23 baseline results by comprehensively understanding the existing implementation and experimenting with various approaches. As will be discussed, we had limited success in improving upon the baseline models.

Our first approach aimed to improve the segmentation baseline of U-Net, which had a validation dice score of 0.823, by utilising nnU-Net. We chose this model due to its ability to automatically configure itself to the context of the dataset including preprocessing, network architecture, training, and post-processing [3]. Further, our training dataset contained 687 central slice 2D training images and as nnU-Net is known for being data-efficient with limited training data, that suggests it could be a good choice. Our implementation scored similarly in validation (**dice score of 0.823**) but performed poorly on the test set (**dice score of 0.43**). One possible reason for this is **overfitting**: though nnU-Net is data-efficient, 687 is a relatively small size for a dataset, and as such, the parameters and architecture could be heavily aligned to the training/validation data. This alignment means that any minor differences in distribution or other variations between the training and test sets could result in the optimised pipeline not generalising well to the test data [5].

Secondly, we replaced the 3D-CNN model for nodule type classification with a Vision Transformer (ViT) applied to the central 2D slice of each 3D nodule CT. ViTs excel at integrating information across entire images via self-attention mechanisms [6], which is useful for our context for capturing texture and relative position of elements in the nodule.

Our validation results showed a **significant decrease in balanced accuracy from 0.855 to 0.622**, indicating the model had not successfully learned the key features necessary to identify the correct nodule type consistently. In Figure 4, the heat map generated by the 3D-CNN shows a more localized and precise activation around the regions that appear to be nodule internal opacity. In contrast, the ViT heatmaps are more diffused and spread out over larger areas of the image, suggesting they might not be as effective in pinpointing small, specific regions that appear to be crucial for nodule type classification. When applied to an isolated nodule, the ViT could not leverage its strength in understanding the broader context or relationships across different regions of the image, as it would in a full lung scan, thus diminishing one of the key advantages of ViTs. Moreover, by selecting the central slice, the ViT also loses the volumetric spatial features that significantly inform the 3D-CNN in this context, indicating e.g the outer shape of the nodule. One limitation of this approach is that the performance of the ViT may have been constrained by the training dataset size of 687. Dosovitskiy et al. (2020) state; "Transformers lack some of the inductive biases inherent to CNNs, such as translation equivariance and locality, and therefore do not generalise well when trained on insufficient amounts of data."

Our final approach involved implementing better data balancing techniques to ensure the classification models were not biased towards the majority classes: the solid nodule class for nodule type classification and the non-malignant nodule class for malignancy classification. This **improved the validation AUC of the 3D-CNN model from 0.821 to 0.855 for nodule type**, but resulted in **poor performance on the test set of 0.508**. Similar issues arose in malignancy classification. After speculating that the cause may be a discrepancy between the training data following augmentation and the test data, we investigated the traits shown in table 2 and found no clear difference. It became clear that **data leakage between the training and validation set** due to data augmentation before doing the train-validation splits was more likely the issue, since it may occur that training and validation sets might contain same images modulus a data augmentation transformation. Therefore, we changed our approach for malignancy classification to ensure the training and validation were totally independent (by first doing the train-validation split, after that performing augmentation in each set separately), resulting in **more reasonable results of 0.967 in validation and 0.862 on the test set**. Due to time constraints we were unable to apply the same retraining to the nodule type classification. These results still fall below the baseline, and future work could involve combining class balancing with hard data mining, augmenting the training dataset with the hardest-to-label images to help the model learn subtle features differentiating nodule types and malignancy levels.

The main limitation of our project was a persistent struggle with **overfitting** across various models, despite incorporating regularization techniques. In some cases, this overfitting can be attributed to the complexity of the models and the small dataset size. Although nnU-Net is data-efficient, 687 scans are still relatively few, leading to potential overfitting. Similarly, ViT's performance may have been constrained by the small training dataset size.

In the wider context of the literature, our work suggests that the **original methods implemented, 3D-CNNs for classification and U-Net for segmentation, remain the strongest models for this context at the moment**.

V. Conclusion

We implemented nnU-Net, ViT, and data augmentation techniques and found that none of these methods improved performance on the test set from the baseline model, likely due to overfitting and unsuitability of the models for this context. Despite these challenges, the insights gained provide a valuable foundation for future research in optimizing model selection and training approaches for lung nodule analysis.

REFERENCES

[1] W. Alakwaa, M. Nassef, and A. Badr, "Lung cancer detection and classification with 3d convolutional neural network (3d-cnn)," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 8, 2017. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2017.080853

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

[3] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[5] D. R. Clymer, J. Long, C. Latona, S. Akhavan, P. LeDuc, and J. Cagan, "Applying machine learning methods toward classification based on small datasets: application to shoulder labral tears," *Journal of Engineering and Science in Medical Diagnostics and Therapy*, vol. 3, no. 1, p. 011004, 2020.

[6] P. Ma, G. Wang, T. Li, H. Zhao, Y. Li, and H. Wang, "Stcs-net: a medical image segmentation network that fully utilizes multi-scale information," *Biomedical Optics Express*, vol. 15, no. 5, pp. 2811–2831, 2024.