Master's Thesis

Analysing the Effect of Multi-Branch Deep Neural Networks on Pulmonary Nodule Classification

A Segmentation-Guided Feature Fusion Approach

Neil Christean Basson

Student number:	15763536
Date of final version:	July 10, 2025
Master's programme:	Data Science and Business Analytics
Specialization:	Business Analytics
Supervisor:	Prof. dr. I. Birbil
Second reader:	Prof. F. Holstege

Faculty of Economics and Business

Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Contributions 2

1.3 Outline 2

2 Literature Review **3**

2.1 Thematic Review of Existing Work 3

2.2 Limitations of Existing Approaches 7

3 Dataset **9**

3.1 Overview 9

3.2 LIDC-IDRI Dataset 10

4 Models **14**

4.1 Data Pipeline 14

4.2 Segmentation 15

4.3 Classification 17

5 Results **21**

5.1 Segmentation Model Performance 21

5.2 Classification Model Performance 24

5.3 Case-Based Analysis of Classification Outcomes 30

6 Conclusion and Future Work **38**

6.1 Summary of Contributions 38

6.2 Key Findings 38

6.3 Limitations 40

6.4 Future Research 40

Appendix **42**

Chapter 1

Introduction

Lung cancer remains one of the most common and deadly forms of cancer globally, accounting for approximately 2.5 million new cases and 1.8 million deaths in 2022 alone, making up 12.4% of all cancer diagnoses and 18.7% of cancer-related deaths worldwide (Ferlay et al., 2024). The prognosis is highly dependent on the stage at which the disease is detected, with five-year survival rates exceeding 60% for localized tumors but dropping below 10% when diagnosed at later stages (American Cancer Society, 2024). Despite the benefits of early detection through screening, many healthcare systems face severe capacity constraints. In the UK, for example, over 17% of CT scans remain unreported after six weeks due to radiologist shortages, which is projected to reach 44% by 2024 (Royal College of Radiologists, 2024). These delays are even more prominent in low-resource settings, where access to timely diagnosis is further limited (Ardila et al., 2019). This context highlights the urgent need for automated tools that can support radiological workflows, reduce diagnostic delays, and improve early detection of lung cancer.

1.1 Motivation

Advances in deep learning have shown strong potential to address these bottlenecks. Recent studies demonstrate that AI-assisted systems can reduce radiologist reading time by up to 27%, with some pre-screening tools achieving reductions of 62% and increasing diagnostic throughput by over 40% (Zhou et al., 2021; Duffy et al., 2023). In clinical trials, AI models have successfully processed low-dose CT scans, ruling out the majority of negative cases (Callister et al., 2024). These developments suggest that integrating AI into the diagnostic workflow could significantly reduce workload, shorten turnaround times, and support large-scale lung cancer screening programs. However, most current models focus on isolated tasks such as segmentation or classification, often ignoring interpretability or pipeline-wide performance. This study addresses that gap by combining segmentation, semantic feature extraction, and interpretable classification within a unified multi-branch neural network (MBNN) framework designed to be efficient, transparent, and clinically relevant. The following research questions were investigated:

1. How does the performance of a MBNN compare to a baseline CNN, where the MBNN includes the baseline CNN as one of its branches?

2. How do the classification decisions of the baseline CNN and MBNN differ in terms of spatial focus, as revealed by Grad-CAM heatmaps?

3. How do semantic feature values relate to model predictions, and how do these relationships differ between correctly and incorrectly classified cases?

4. How does the performance of the full pipeline compare to models that handle classification or segmentation in isolation, and how does the classification accuracy of the MBNN change when using predicted segmentation masks instead of ground-truth masks as model input?

5. How do the tumour mask shapes and the semantic feature values extracted from them differ between predicted segmentation outputs and radiologist-provided ground-truth masks?

1.2 Contributions

This study addresses these challenges by developing a fully integrated deep learning pipeline for pulmonary nodule analysis, covering the entire diagnostic flow from segmentation to classification. The pipeline first applies a U-Net++ model to localize the tumor region, followed by semantic feature extraction from the resulting binary mask. These features are then combined with the original CT crop and the mask itself within a MBNN. In addition to predicting malignancy, the network incorporates semantic descriptors and explainability tools to better understand the model's reasoning. Unlike prior studies that isolate segmentation or classification or focus solely on performance, this work explores the full interaction between components and examines which tumor characteristics most strongly affect prediction. Notably, the entire process, from raw CT input to malignancy prediction, is fully automated, removing dependence on radiologist-drawn masks or hand-crafted semantic annotations. This supports scalable, interpretable, and resource-efficient screening workflows that are better suited for real-world clinical deployment.

1.3 Outline

The remainder of this study is structured as follows. Chapter 2 provides a review of existing literature covering deep learning methods for lung nodule segmentation, classification, and multi-branch (MB) network architectures. Chapter 3 outlines the data sources and preprocessing steps used throughout the pipeline. Chapter 4 describes the model architectures and implementation details for both segmentation and classification tasks. Chapter 5 presents the evaluation results and interpretation of model performance. Finally, Chapter 6 discusses the findings, limitations, and potential directions for future work.

Chapter 2

Literature Review

This chapter provides a structured review of recent deep learning methods applied to pulmonary nodule analysis in CT imaging. It first examines segmentation models used to localize nodules, followed by an overview of classification approaches for malignancy prediction. A subsequent section explores MBNN architectures that integrate multiple data modalities to enhance classification performance. The chapter concludes by outlining key limitations and gaps in the existing literature that form the motivation and design of this study.

2.1 Thematic Review of Existing Work

Segmentation Models

Table 2.1 presents a selection of existing segmentation models along with their year of publication, architecture, input size, and reported Dice Similarity Coefficient (DSC). These studies demonstrate a range of architectural approaches, from early models with large input sizes to more modern and specialized designs. The CF-CNN model from 2017 uses a 572×572 input size to support coarse feature learning across the full spatial context of the CT slices (S. Wang et al., 2017). The DCNN model introduced in 2019 operates on 96×96 nodule-centered patches and achieves the second strongest segmentation performance of 0.8310 DSC (Tang et al., 2019). The U-Net model from 2021 works with 64×64 cropped patches and achieves relatively high segmentation performance, reflecting the strength of its encoder-decoder structure (Kumar et al., 2021). A 3D GAN-based architecture proposed in 2022 segments 64×64×32 voxel blocks, illustrating a shift toward volumetric segmentation and generative modeling techniques (Tyagi & Talbar, 2022).

Finally, the U-Net++ model from 2023 uses 128×128 input slices and dense skip connections to enhance model performance and it reported a DSC of 0.8670, the highest of all listed methods (Lin et al., 2023). These models collectively reflect the progression and variety in segmentation strategies, from traditional CNNs to more complex 3D and nested architectures.

Reference	Year	Architecture	Input Size	Dice Similarity Score
(S. Wang et al., 2017)	2017	CF-CNN	572x572	0.8214 ±0.1076
(Tang et al., 2019)	2019	DCNN	96x96	0.8310 ±0.0885
(Kumar et al., 2021)	2021	UNet	64x64	0.8205
(Tyagi & Talbar, 2022)	2022	3D GAN	64x64x32	0.8074
(Lin et al., 2023)	2023	UNet++	128x128	0.8670 ±0.0077

Table 2.1: Performance of representative lung-nodule segmentation models reported in the literature

This study focuses on the U-Net and U-Net++ architectures, which are well suited for biomedical image segmentation tasks due to their encoder decoder design and capacity to capture multi-scale features. These features range from low level edges and textures to high level abstract representations, enabling the network to segment structures with varying sizes and complexity more effectively. The U-Net model of 2021 was trained on 2601 nodules containing CT slices from the LIDC-IDRI dataset, with each slice cropped to 64×64 around the tumor. It achieved an accuracy of 0.9445 and a DSC of 0.8205 using the Adam optimizer (Kumar et al., 2021). In comparison, the U-Net++ model using 2D slices resized to 128×128 also used the LIDC-IDRI dataset and reported an accuracy of 0.9265 and a DSC of 0.8670 (Lin et al., 2023). These results provide strong benchmark performance for both models and support the effectiveness of 2D segmentation approaches on the LIDC-IDRI dataset. These two architectures were chosen based on their consistent performance in prior studies, their proven suitability for medical image segmentation, and their compatibility with the resolution and structure of the available LIDC-IDRI data.

Classification Models

After the segmentation step, the final part of the cancer detection pipeline is to classify the nodule as either malignant or benign. Most recent studies reduce the LIDC-IDRI malignancy ratings (1–5) to binary labels, but labeling strategies vary. Some exclude nodules rated 3 as ambiguous (Causey et al., 2018), while others group 1–3 as benign and 4–5 as malignant (Ali et al., 2020). This inconsistency complicates direct comparison between models and is why the models excluding malignancy ratings of 3 are further investigated below Table 2.2, as this aligns closer to what is implemented in this study.

Table 2.2 below shows a summary of some of the main studies found. All of them use convolutional nerual networks (CNN's), 3D CNNs, or transformer-based models, with input sizes typically around 64×64. This makes sense since most studies crop around the region of interest, which is the segmented tumor area. Accuracy scores range from about 0.8900 to nearly 0.9700. For example, (Ren et al., 2020) reported 0.9226 using a basic 3D CNN, while (Wu et al., 2022) got 0.9470 with a transformer model.

Some older models like those from (Causey et al., 2018) combined a 3D CNN with a random forest classifier, and still achieved strong results (above 0.9300). It shows that even shallow architectures can work well when combined with ensemble techniques.

Reference	Year	Architecture	Input Size	Accuracy
(Nibali et al., 2017)	2017	ResNet	64x64	0.9107
(Causey et al., 2018)	2018	3D CNN + Random Forest	64x64	0.9370
(Xu et al., 2020)	2020	3D CNN	64x64x64	0.9445
(Ali et al., 2020)	2020	2D CNN	48x48	0.9265
(Ren et al., 2020)	2020	3D CNN	32x32x32	0.9226
(Abid et al., 2021)	2021	2D CNN	40x40	0.8930
(Wu et al., 2022)	2022	Transformers	64x64	0.9470
(Zhang et al., 2019)	2023	Transformers	64x64	0.9236

Table 2.2: Performance of representative lung-nodule classification models reported in the literature

The 2 main studies used for benchmarking are analysed below:
The study by (Causey et al., 2018) presents a 3D CNN model for lung nodule malignancy prediction using CT scans from the LIDC-IDRI dataset. They evaluated two setups: one using only nodules rated 1 and 5 (achieving 0.9620 accuracy), and another including ratings 1–2 as benign and 4–5 as malignant (0.9370 accuracy). The model processes ten CT slices per input and uses five convolutional layers followed by two dense layers. While the number of trainable parameters isn't stated, the study demonstrates that high accuracy is achievable with limited but carefully selected input.

The study by (Xu et al., 2020) developed a 3D CNN with residual connections to classify lung nodules. The model takes 64×64×64 voxel cubes as input and includes four residual blocks, followed by global average pooling and fully connected layers. To avoid label ambiguity, they only used nodules rated 1 or 5 for training and testing. The model achieved an accuracy of 0.9445, with a sensitivity of 0.9545, specificity of 0.9345, and an AUC of 0.976.

Multi-Branch Models

Recent work has increasingly explored the use of MBNNs in pulmonary nodule classification. These architectures offer a flexible way to combine different types of input such as CT image patches, segmentation masks, semantic labels, or hand-crafted features, into a single unified model. By processing each modality in a dedicated branch, MBNNs allow networks to extract complementary information that might be lost in single-input designs. Several recent studies have adopted this structure to investigate how performance and interpretability can be improved using multimodal inputs.

The three peer-reviewed studies below are used as benchmarks for performance comparison. Each paper proposes a different MBNN configuration and focuses on a specific aspect of the classification task, such as semantic grading, multi-modal fusion, or visual attention. These models form the reference point for later comparisons and help identify literature gaps addressed in this research. The reported performance metrics for these models are summarised below in Table 2.3

Reference	**Year**	**Model**	**Accuracy**	**AUC**
(Liu et al., 2023)	2023	MBCNN semantic grading	0.8937	–
(Yuan et al., 2023)	2023	3D ECA-ResNet + semantic fusion	0.9489	0.9784
(Zheng et al., 2024)	2024	Multichannel attention CNN	0.9011	0.9566

Table 2.3: Reported performance metrics of selected multi-branch neural networks for lung nodule classification

The first study by (Liu et al., 2023) focuses on predicting multiple radiologist-defined semantic characteristics of lung nodules rather than solely classifying malignancy. It introduces a MBCNN designed to grade six semantic traits, using three distinct views of each nodule as input. These include a 64 by 64 grayscale CT crop, its corresponding binary mask, and a third stream that fuses skip-connected features to enhance representation. Each branch passes through a series of four convolutional layers before being merged and forwarded to six parallel output heads, each corresponding to one semantic feature. The network is trained on 3 091 nodule slices extracted from 200 LIDC IDRI scans, using a binary label for each trait based on the radiologists' original scores. The architecture achieves strong performance across all six targets, outperforming both standard CNN and Deep Belief Network baselines. The study does not apply post hoc explainability methods such as Grad CAM, nor does it analyse how the predicted semantic grades contribute to malignancy classification. Although this study does not investigate malignancy prediction, it serves as a relevant reference by showcasing how MBNNs can be applied in lung cancer-related deep learning tasks.

The second study by (Yuan et al., 2023) investigates whether combining image and structured semantic information improves classification of pulmonary nodules. The proposed multi-branch architecture consists of two inputs: a 64 by 64 by 64 voxel cube centered on the nodule and a nine-dimensional vector of clinical and radiological attributes. The image branch uses a 3D ECA-ResNet model with sub-branches to capture features at different depths. The semantic branch is a multi-layer perceptron (MLP) that receives features manually extracted by radiologists from LIDC-IDRI annotations, including diameter, subtlety, internal structure, calcification, sphericity, margin, lobulation, spiculation, and texture. The model is trained on a subset of LIDC-IDRI filtered through the LUNA16 protocol, with 686 nodules labeled as benign or malignant, based on an average malignancy score: ratings of 1–3 are labeled benign, while scores

of 4–5 are considered malignant. Grad CAM is applied to the image branch to verify that the network attends to the relevant nodule regions, offering basic interpretability. The main contribution is demonstrating that multi-modal fusion of volumetric image data with structured semantic features improves classification accuracy compared to single input models. However, the study does not analyze how individual features contribute to predictions or compare interpretability and performance across model variants.

The third study by (Zheng et al., 2024) proposes a MBNN that integrates three complementary image representations of lung nodules to improve malignancy classification. The model consists of three parallel branches, each processing a different modality derived from the same axial CT slice: the original grayscale image, a binary segmentation mask, and a Local Binary Pattern (LBP) texture map. Each branch is implemented as a convolutional subnetwork with identical architecture, and their outputs are fused at a late stage using a spatial attention module. All inputs are 64 by 64 pixel patches centered on nodules from the LIDC-IDRI dataset. The model is trained using malignancy scores provided by four radiologists per nodule, with binary labels assigned by majority vote: nodules with an average score above 3 are classified as malignant. The attention mechanism improves performance by learning spatial weighting across branches, but no explainability techniques such as GradCAM are applied to interpret how the network combines the three input types. This limits the interpretability of the model's decisions and leaves open questions about what visual characteristics drive the classification.

2.2 Limitations of Existing Approaches

While segmentation, classification, and MB designs have each been extensively studied, current literature tends to treat these components in isolation. Segmentation models focus on localizing the tumor accurately, while classification models, whether 2D, 3D, or transformer-based, optimize predictive performance using preprocessed inputs. The reviewed MBNNs further demonstrate that combining multiple data modalities improves classification accuracy and potentially interpretability. However, none of the studies develop and evaluate the full cancer detection pipeline from beginning to end.

For example, (Lin et al., 2023) approach semantic analysis as a standalone supervised task, predicting radiologist-defined traits independently from malignancy. (Yuan et al., 2023) show that fusing 3D CT volumes with manually extracted semantic features improves accuracy, but their study does not explore where these features originate from or how their interaction with image-based features contributes to final predictions. (Zheng et al., 2024) apply attention-based fusion of image modalities but exclude semantic or clinical information entirely, and provide no GradCAM style interpretability of the combined representations. Across all three MBNN's, semantic features are hand labelled from LIDC-IDRI metadata rather than extracted from

segmentation outputs, and explainability remains limited to visual heatmaps or is omitted altogether.

This study addresses these limitations by integrating the full imaging pipeline. A UNet++ model first segments the CT scan and produces a binary tumor mask. From this, four semantic features, spiculation proxy, compactness, solidity, and diameter, are automatically extracted and used as one of three inputs to the final classification model. The other two branches include a cropped grayscale CT slice and the corresponding binary tumor mask. The resulting MBNN is designed not just for performance, but for analysis. The structure allows investigation into how each modality contributes to the model's predictions and how interpretability tools can reveal decision-relevant patterns. By utilizing the segmented masks as the input for the MBNN classification model, it allows us to analyze the impact the use of such masks have on classification performance, as opposed to using ground truth tumor masks. In doing so, this study builds a full pipeline framework that reflects more realistic diagnostic workflows while exploring under-researched areas like feature interaction and modality-specific influence.

Chapter 3

Dataset

This chapter outlines datasets that are currently publically available and more specifically the dataset used for training and evaluating the models. This chapter then explains how the data was filtered, preprocessed, and split for both segmentation and classification tasks.

3.1 Overview

Among the available datasets, LIDC-IDRI stands out as the foundation for much of the existing work in lung cancer detection. It's known not just for the size of the dataset, but because each scan is annotated independently by four radiologists, enabling analysis of differences between radiologist annotations or building ensemble ground truths using strategies like majority voting (Armato III et al., 2011). LUNA16 is a curated subset of LIDC-IDRI that filters out scans with thicker slices and structures the data specifically for developing and benchmarking nodule detection algorithms. It's often used when the goal is to reduce false positives in detection pipelines (Setio et al., 2017). NSCLC-Radiomics offers something a bit different, it focuses on patients with non-small cell lung cancer and includes segmentation masks of the tumors along with clinical outcome data. That makes it especially valuable for feature extraction tasks like radiomics or for predictive modeling based on pre-treatment scans (Aerts et al., 2019). Lastly, RIDER Lung PET-CT is mainly used when researchers want to combine functional and anatomical imaging. Because it includes paired PET and CT scans, it supports studies that focus on treatment response, multimodal registration, or even image synthesis between modalities (P. Wang et al., 2023). A breakdown of the contents of each dataset can be seen below in Table 3.1.

Name	Scope	File Type	Size
LIDC-IDRI	1'010 patients with 244'527 slices across 1308 CT series	DICOM	133GB
LUNA16	888 CT series	MetaImage	65GB
NSCLC-Radiomics	422 patients with 52'073 slices across 1'265 CT series	DICOM & RTSTRUCT	35.8GB
RIDER Lung PET-CT	244 patients with 266'280 slices across 1328 series	DICOM	83GB

Table 3.1: Overview of publicly available lung imaging datasets for pulmonary nodule analysis

3.2 LIDC-IDRI Dataset

LIDC-IDRI is by far the most commonly used dataset for developing and validating machine learning models in lung cancer research (Armato III et al., 2011). What sets it apart is the level of detail and consistency across its annotations, along with the fact that it was built collaboratively by multiple research institutions. The dataset includes rich metadata and standardized scan formats, which simplifies preprocessing and downstream analysis. Its broad adoption has led to the development of several benchmark datasets built directly from it, including LUNA16. This makes it a common reference in studies looking to compare model performance. The consistency of its structure and the fact that it's fully open-access has made it a foundational resource in lung CT research.

Annotation Protocol

Each CT scan in the LIDC-IDRI dataset is reviewed by four board-certified radiologists, all working independently. The 3D scan is represented as a series of 2D axial slices, and the radiologists annotate nodules with a diameter of 3 mm or more by outlining them on the individual slices where the nodule appears. Since no discussion or coordination happens between the annotators, the result is a set of masks that naturally capture annotation variability among readers. To reduce the impact of these differences, a consensus mask is created by combining the annotations: a pixel is included in the final ground truth mask if at least two out of the four radiologists marked it as part of the nodule. This simple majority rule makes the resulting mask more robust and reduces the influence of any single outlier opinion. It also reflects a more realistic representation of how nodules are perceived in clinical settings, where interpretation can vary between observers.

Preprocessing and Data Handling

Before training the segmentation model, each 2D CT slice undergoes a series of focused preprocessing steps, which can be seen in Figure 3.1. The first is intensity standardisation, where pixel values are rescaled to have a mean of zero and a standard deviation of one. This helps

reduce differences caused by varying imaging equipment or acquisition settings and makes the training data more uniform. To deal with noise in the images, a combination of median filtering and anisotropic diffusion is applied. The median filter reduces salt-and-pepper noise while preserving edges, and the anisotropic diffusion smooths out uniform areas while keeping structural boundaries sharp.

Next, the lungs are extracted using an adaptive thresholding approach based on K-Means clustering. By separating the pixel intensities into two groups, roughly corresponding to the lungs and the background, a threshold is selected that enables the creation of a binary mask. Morphological operations such as erosion and dilation are used to clean this mask, removing small specks and filling in gaps. The two largest connected components are then selected, as these usually correspond to the left and right lung fields. This binary mask is applied to the original image to isolate the lung regions. Everything outside of the selected regions is set to zero, so only the lung tissue remains.

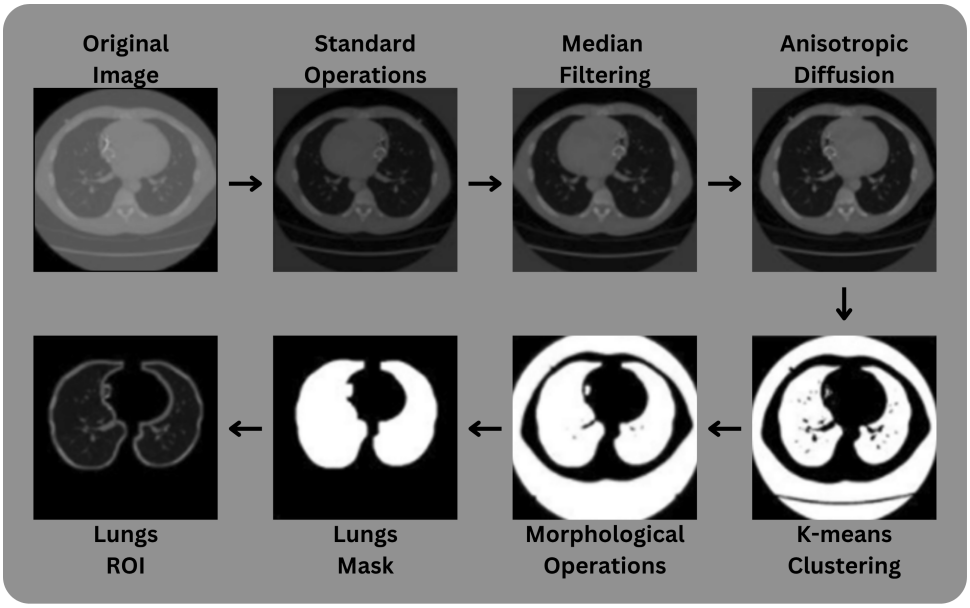

Figure 3.1: Sequential visualisation of the preprocessing pipeline applied to CT slices

Segmentation Subset

For training the segmentation model, a subset of 500 patients was selected from the full LIDC-IDRI dataset, which contains thoracic CT scans from 1,018 patients. Each scan consists of hundreds of axial slices, resulting in over 244,000 individual 2D grayscale images at a resolution of 512×512 pixels. However, not all slices contain visible pulmonary nodules. After filtering, only approximately 9,000 slices were retained for use in the segmentation task. This reduction comes from two main factors: (1) the spatial sparsity of nodules, which typically occupy a small fraction of the axial slices in each scan, and (2) the need to exclude ambiguous or inconsistent annotations. Of the 9,000 selected slices, about 7,000 contained one or more clearly visible nodules with corresponding binary masks, while the remaining 2,000 were nodule-free

slices randomly sampled from the same patient set. These negative examples were deliberately included to help the model learn when not to segment, enhancing its discriminative capacity and reducing false positives. The final dataset was partitioned into training, validation, and test sets using an 80/10/10 split to maintain balanced representation across subsets.

Classification Subset

For the classification model, preprocessing begins by localizing the tumor using the binary ground truth mask. The centroid of each tumor is computed from the mask and used to define the central coordinates for cropping a region of interest (ROI) from the corresponding CT slice. A square window of 64×64 pixels is then extracted around this center, capturing the nodule and its immediate context. This approach ensures that the cropped image consistently centers the tumor, even for nodules with irregular shapes. These 64×64 grayscale ROIs form the input for the image branch of the classification model. Approximately 2,000 of these tumor-centered crops were selected to train a baseline CNN.

In parallel, the binary tumor mask is also cropped using the same center coordinates, resulting in a 64 by 64 binary mask aligned with the original ROI. This mask is not used in the baseline model, but it plays a key role in the MBNN, where it serves as a separate input channel. From the shape of the tumor in this binary mask, four semantic features are extracted: spiculation proxy, compactness , diameter, and solidity. These features provide an abstract but interpretable description of the tumor's geometry and are used as an additional MLP input branch in the MBNN model.

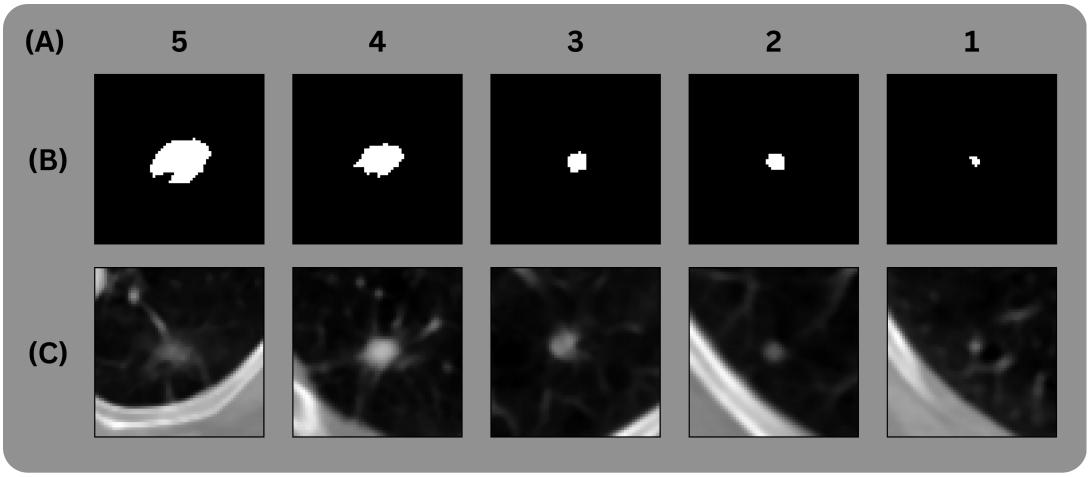

Figure 3.2: Visual comparison of pulmonary nodules across malignancy ratings with corresponding masks and CT regions

Each nodule in the dataset has been scored for malignancy by multiple radiologists on a scale from 1 to 5. To obtain a single label for each sample, the median of the available ratings is used. Figure 3.2 presents representative examples of nodules across all five malignancy scores. For each example, row A indicates the assigned malignancy score, row B shows the corresponding

cropped 64x64 binary tumor mask, and row C displays the cropped 64×64 greyscale CT slice from which the mask was derived. Nodules with a median score of 1 or 2 are labeled as benign, while those with a score of 4 or 5 are labeled as malignant. Malignancy score 3 was excluded due to its ambiguous clinical interpretation. It represents nodules for which radiologists could not confidently assign a benign or malignant label. Including such uncertain cases could introduce noise and degrade model performance by blurring the distinction between classes. To ensure a more reliable ground truth for training and evaluation, only nodules with ratings of 1–2 (clearly benign) and 4–5 (clearly malignant) were kept.

Chapter 4

Models

Understanding the structural design of deep learning models is essential to both segmentation and classification tasks within the proposed lung cancer detection pipeline. This chapter presents a detailed breakdown of the architectures developed for each stage of the system. Specifically, this chapter describes the segmentation model used to delineate tumor regions from CT scans and the classification models responsible for identifying whether a pulmonary nodule is cancerous or not. These models serve as core components of the pipeline and are critical to its overall accuracy and robustness.

4.1 Data Pipeline

The complete data pipeline used in this study is illustrated in Figure 4.1. The process begins with the acquisition of a CT scan, where the patient undergoes imaging and the resulting DICOM file is generated. This raw CT data is then passed through a preprocessing stage, where it is normalized, resized, and prepared for model input. The preprocessed 512×512 greyscale CT slices are subsequently fed into the segmentation model, which produces a binary mask outlining the tumor region. From this segmentation output, a 64×64 region of interest is cropped from both the original CT scan and the segmented binary tumor mask. In addition, four semantic shape features are extracted from this binary mask. These three data streams (cropped CT image, cropped mask, and semantic features) are then used as inputs to the MBNN classification model, which outputs a binary prediction indicating whether the tumor is malignant or benign.

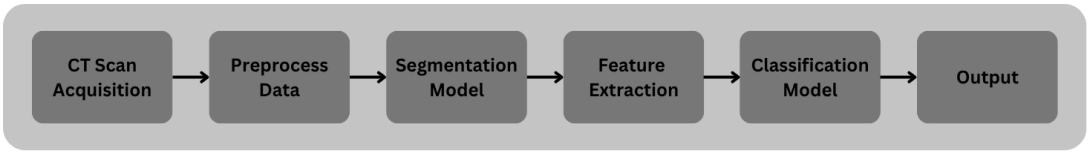

Figure 4.1: Overview of the end-to-end pipeline for pulmonary nodule analysis

4.2 Segmentation

Segmentation serves as the initial step in the image analysis workflow, focusing on the delineation of tumor regions from lung CT scans. The primary objective is to produce binary masks that clearly separate tumor structures from surrounding tissue, allowing for both visual assessment and the extraction of meaningful features for classification. Segmentation helps identify the tumor location within the CT scan, enabling the extraction and cropping of the region of interest that is later used during classification.

UNet++

To carry out the segmentation, a U-Net++ model was used. This architecture builds on the original U-Net but adds more connections between the encoder and decoder layers. These extra paths allow the model to better combine low-level and high-level features at different resolutions. As seen in the diagram Figure 4.2, the architecture includes multiple intermediate nodes between each encoder and decoder level, which helps to reduce the gap in feature representations across the network. This kind of design tends to perform well in medical image segmentation, especially in cases like this where precise boundaries and local context matter a lot.

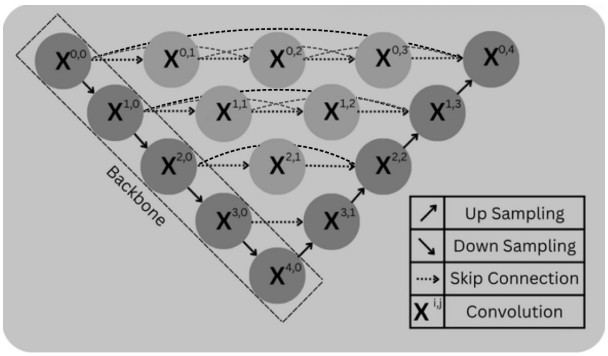

Figure 4.2: Nested skip-connection scheme in the UNet++ encoder–decoder

Figure 4.3 shows the standard U-Net backbone on which U-Net++ is built. The encoder contains five stages; each stage applies two 3×3 convolutions followed by 2×2 max-pooling, doubling the channel depth from 64 up to 1 024. The decoder mirrors this pattern with bilinear up-sampling and concatenation-based skip connections, bringing the feature maps back to the original 512×512 resolution.

U-Net++ extends this backbone by inserting intermediate convolutional blocks between every encoder and decoder level (see Figure 4.2). These densely connected paths fuse low, mid and high-level features, improving boundary localisation, an advantage for segmenting small or non-distinct nodules in CT scans. Each intermediate block consists of two convolutional layers with batch normalisation and ReLU activation, and a final 1×1 convolution converts the decoder output to a single-channel tumour-versus-background probability map. The architecture with filter counts and output sizes can be seen in Table 7.1 in the Appendix.

The complete network contains about 9 million trainable parameters, reflecting the added depth and dense connectivity without a prohibitive increase in model size.

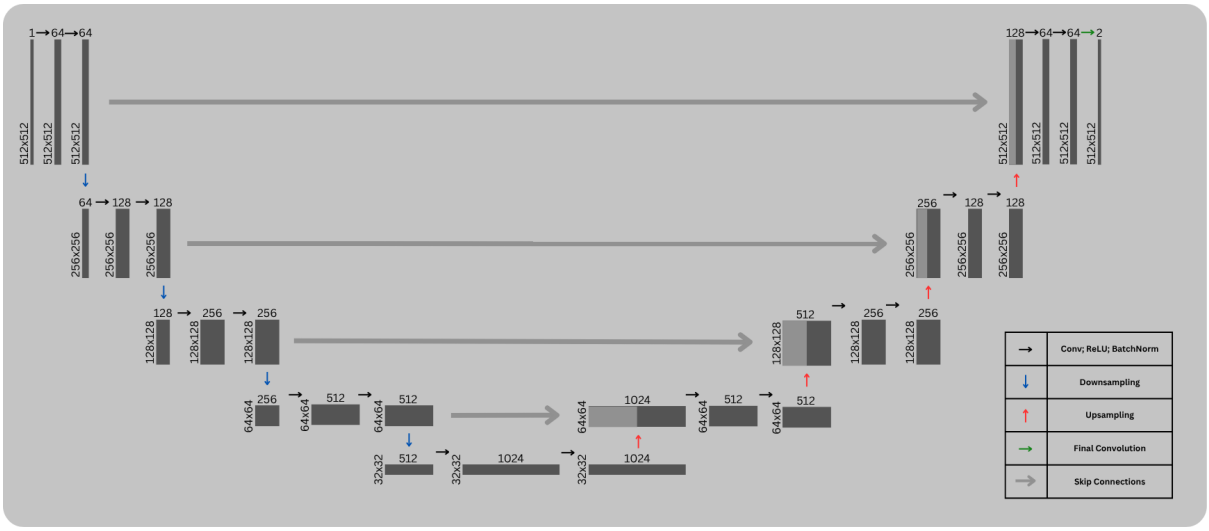

Figure 4.3: Layer-wise layout of the implemented UNet segmentation network

Model Training

The segmentation model was trained using a combination of binary cross-entropy (BCE) loss and Dice loss, which are commonly used for medical image segmentation tasks. The BCE component ensures pixel-level accuracy, while the Dice loss helps handle class imbalance by focusing more on the overlap between predicted and ground truth masks. In addition to this, Intersection over Union (IoU) was also tracked during training as a supplementary evaluation metric, since it provides a good indication of how well the predicted masks match the actual tumor regions.

For optimization, the Adam optimizer was used with a learning rate of 1×10^{-5} and a weight decay of 1×10^{-4}. These values were chosen after conducting hyperparameter testing across different learning rates and regularization strengths. This configuration led to the most stable convergence and best overall segmentation performance. The model was trained for 100 epochs using a batch size of 8, which was also selected based on empirical testing. Other batch sizes and training lengths were explored, but this setup consistently resulted in better generalization and smoother training curves.

To support the computational demands of training the U-Net++ segmentation model on high-resolution 512×512 input slices, Google Colab Pro was used with access to an NVIDIA A100 GPU. This high-memory, high-throughput GPU enabled batch processing and efficient back-propagation for the deep encoder-decoder architecture. Despite this, the full training cycle still required approximately 10 to 11 hours due to the model's complexity, the large input size, and the use of combined loss functions. Leveraging Colab Pro's extended session durations and

GPU resources was therefore essential to making training computationally feasible within a single workflow.

Integration into the Classification Pipeline

The segmentation model plays an important role in the overall workflow by generating binary masks that accurately outline the tumor region in each CT slice. These masks are not only useful for visualization but are also essential for computing the semantic features used later in the classification stage.

The semantic feature values like compactness, solidity, diameter, and a spiculation proxy are all derived directly from the output of this model. Because of this, the quality of the segmentation has a direct impact on the quality of these features. If the segmentation is inaccurate or inconsistent, the extracted features can be distorted or biased, which would negatively affect the classification performance. In that sense, the segmentation model forms a critical foundation for the next steps in the pipeline, acting as the link between raw imaging data and higher-level feature-based analysis.

4.3 Classification

Once segmentation has been performed, the classification stage evaluates the identified tumor regions to determine whether they are malignant or benign. This stage consists of two model designs: a baseline CNN and a more sophisticated MBNN. Each model leverages different types of input data and abstraction levels, with the baseline CNN providing a reference point and the MBNN integrating diverse features for improved decision-making. This section provides architectural details and training considerations for both models.

Baseline Convolutional Neural Network (CNN)

This section introduces the baseline convolutional neural network (CNN) developed for the binary classification of lung tumors present in CT scan images. The model is designed to be structurally simple and computationally efficient, with the objective of serving as a foundational reference point to later be used for benchmarking the performance of more complex architectures. It is also constructed to enable later integration into a MBNN architecture without introducing excessive parameter complexity and ensuring computational efficiency.

The baseline CNN processes single-channel grayscale input images of size 64×64 pixels, each representing a ROI centred around a tumor. These ROIs are identified in the original CT slices and this procedure is explained in Chapter 3. This localized input allows the model to focus exclusively on the most relevant areas for classification. By constraining the spatial context, the model is able to avoid irrelevant anatomical structures and drastically decrease the input dimensionality.

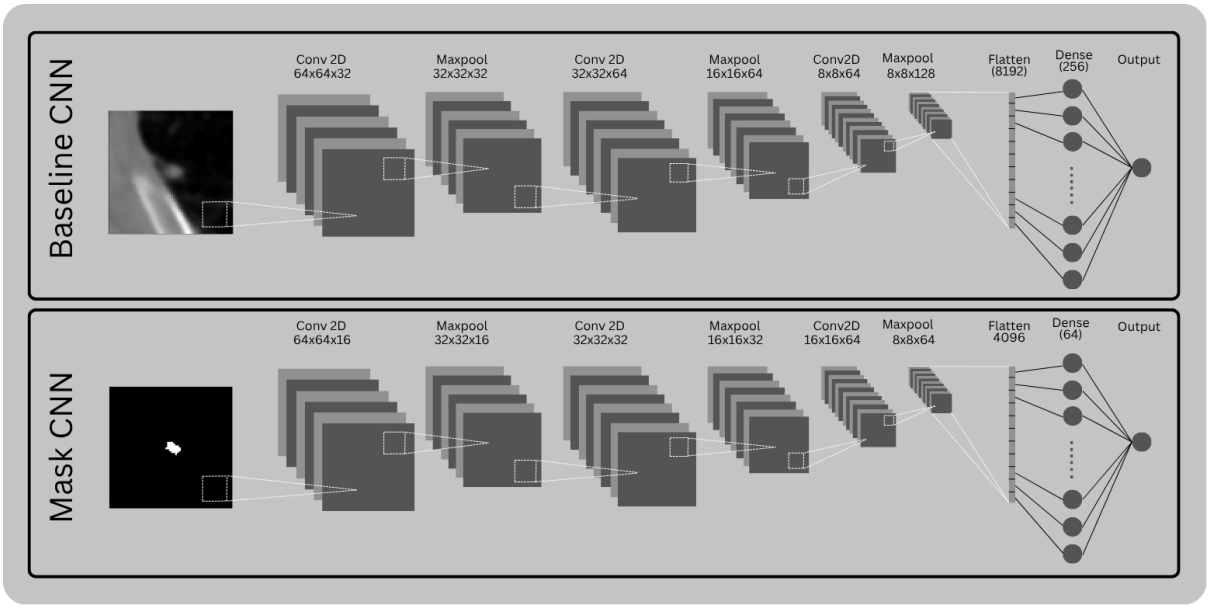

Figure 4.4: Comparison of the baseline CNN and mask-based CNN branch used in the multi-branch network

The model architectures, displayed in Figure 4.4, follows a standard feed-forward convolutional structure comprised of three convolutional blocks, each followed by max pooling and a fully connected output layer. Each convolutional layer uses a 3×3 kernel with padding to preserve spatial dimensions and applies the ReLU activation function. A dropout layer is inserted before the final classification layer to reduce overfitting risk during training. The model is trained using the Adam optimizer, and the binary cross-entropy loss function is used to optimize performance for the binary classification task. The total number of trainable parameters in the model is approximately 2.19 million, which allows the baseline model to learn meaningful representations.

Model Training

Hyperparameter testing was conducted to identify the optimal configuration for training the baseline model. This involved evaluating the effects of different batch sizes and learning rates on both training dynamics and classification performance. Batch sizes of 8, 16, 32, and 64 were tested. While all configurations achieved convergence, the model trained with a batch size of 16 produced the most consistent and well-balanced results. It outperformed the others across key evaluation metrics, including accuracy, precision, recall, specificity, and F1 score. Although the differences between configurations were in some cases marginal, batch size 16 demonstrated the most stable learning behavior and offered the best overall trade-off between sensitivity and false negative control.

Multi-Branch Neural Network (MBNN)

This section introduces the proposed three-branch neural network developed for classification.

The model is designed to combine different but complementary types of information extracted from the same tumor region. It consists of three parallel branches, each processing a distinct input modality. These branches extract features independently before their outputs are merged through a concatenation step. The combined feature representation is then passed through a fully connected layer and a final output unit to perform binary classification.

The first branch corresponds to the baseline CNN described in the previous section and processes the 64x64 greyscale ROI extracted from the CT scan.

The second branch is a CNN that processes a 64×64 binary segmentation mask of the tumor. Its architecture can be seen in the bottom half of Figure 4.4. Since this input contains only the binary spatial outline, without texture or intensity information, the branch uses a simpler architecture with fewer filters and a smaller dense layer compared to the image branch. Its primary role is to extract structural and shape-related features from the tumor region, which may contribute to the overall classification accuracy.

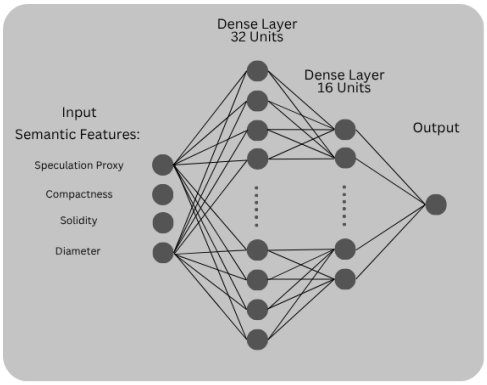

Figure 4.5: Architecture of the semantic feature branch in the multi-branch neural network

The third branch is a multi-layer perceptron (MLP) that processes a four-dimensional vector of semantic features extracted from the tumor's binary mask. Its architecture can be seen in Figure 4.5 These features quantify geometric and morphological properties of the tumor, such as diameter, solidity, compactness, and a spiculation proxy. This branch is fully connected and helps to introduce high-level, hand-crafted descriptors into the model's decision-making process.

Although each branch processes a distinct input modality through its own subnetwork, the branches are not fully independent. They operate in parallel on different data representations, allowing each to extract complementary feature sets. These outputs are then flattened and concatenated into a unified feature vector, which is passed through a shared dense layer followed by a sigmoid-activated output unit that produces a probability score for binary classification. The model uses the same loss function and optimizer as the baseline CNN. All branches are trained simultaneously, with gradients propagated through the entire network. A complete ar-

chitectural diagram is provided in Figure 7.1 in the Appendix.

This design performs feature-fusion, allowing the model to learn interactions across different modalities, as opposed to averaging separate predictions from isolated models. This fusion enables the network to integrate both low-level spatial features and high-level semantic information into a comprehensive decision-making process.

A complete layer-by-layer architecture with output dimensions and activation functions is provided in Table 7.2. The total number of trainable parameters in the model is approximately 2.25 million, which is only marginally higher than the baseline CNN, having 2.19 million trainable parameters. This relatively small increase in computational complexity is an important consideration when comparing model performance. If later evaluation demonstrates a significant improvement in classification accuracy, it would support the use of multi-branch networks as an effective and computationally feasible improvement over single-branch models, particularly in the context of medical image analysis for cancer classification, where both accuracy and efficiency are critical.

Model Training

Hyperparameter testing was conducted to determine the optimal batch size for training the MBNN. Batch sizes of 16, 32, 64, and 128 were tested, with each configuration once again evaluated in terms of training dynamics and classification performance. The same key metrics as before were considered.

Of all the tested configurations, batch size 64 delivered the most favorable overall performance. It provided the best balance across core metrics, demonstrating strong precision and specificity while maintaining high recall. Compared to batch size 16, which achieved slightly higher recall, the batch size 64 produced fewer false negatives and showed greater overall stability. Batch size 32 performed reasonably well but exhibited more variability in validation performance, with signs of reduced generalization. Batch size 128 was the least effective, showing signs of underfitting and slower convergence, especially in later epochs. Based on this, a batch size of 64 was selected for training the MBNN.

Chapter 5

Results

This chapter presents the empirical findings of the study, beginning with the performance of the UNet++ segmentation model and proceeding to an evaluation of MBNN and baseline CNN for malignancy classification. Section 5.1 reports pixel-level metrics, examples of predicted masks, a comparison with published work, and an analysis of mask drift across malignancy ratings. Section 5.2 examines the classification results through confusion matrices, precision-recall statistics, and receiver-operating-characteristic curves, followed by a benchmark against recent studies listed in the literature review. Section 5.3 provides a case-based analysis that combines Grad-CAM visualisation with semantic-feature statistics to identify factors influencing correct and incorrect predictions. Together, these results quantify the contribution of accurate segmentation and multimodal feature fusion to reliable lung-nodule diagnosis.

5.1 Segmentation Model Performance

The segmentation model demonstrated strong overall performance across multiple evaluation metrics on the test set. It achieved an Intersection over Union (IoU) of 0.8038, a DSC of 0.8846, accuracy of 0.9998, and precision of 0.8670. These results indicate that the model was able to accurately detect and segment pulmonary nodules, with high spatial agreement between predicted and ground truth masks. The high DSC and IoU values reflect effective overlap with ground truth, while the near-perfect accuracy suggests that most background regions were correctly ignored. Precision, while slightly lower, highlights the model's conservative approach in avoiding false positives.

Examples of segmentation outputs on the test set and their corresponding ground truth masks are shown in Figure 5.1. Top row (A) indicates the malignancy rating. Row (B) displays the original CT slices, row (C) shows the corresponding ground truth masks, and row (D) presents the predicted segmentation outputs. The results illustrate the model's ability to segment nodules with varying malignancy severity.

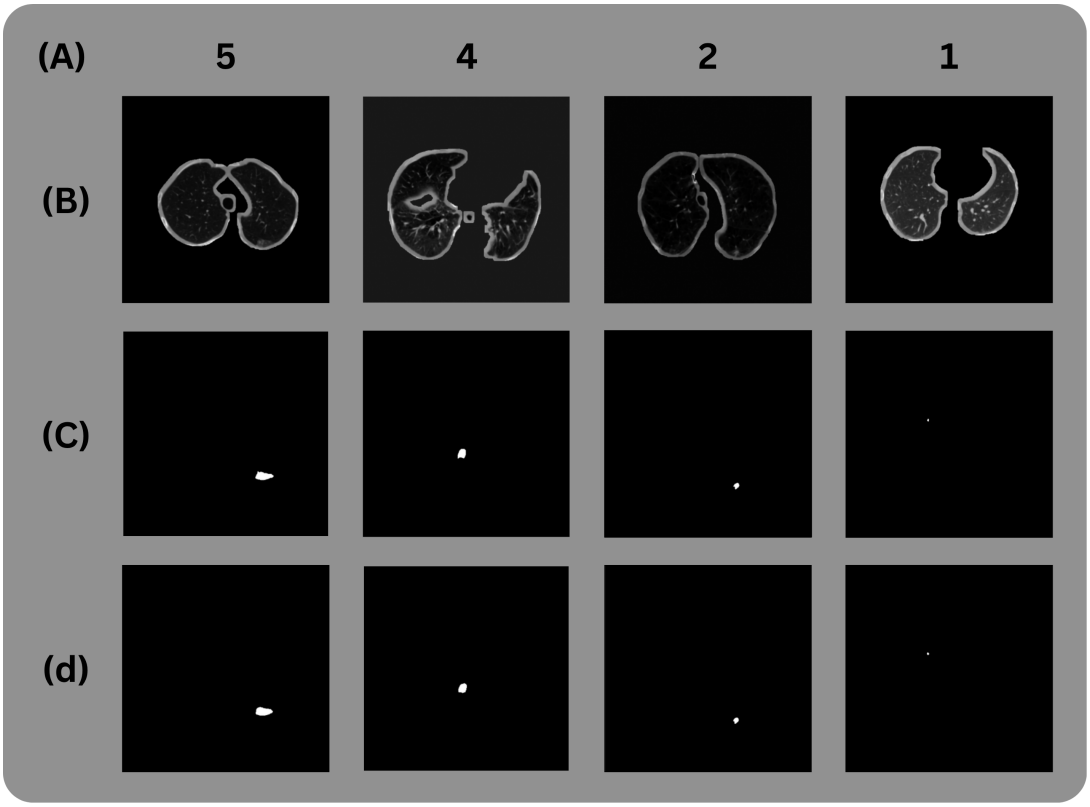

Figure 5.1: Segmentation output across malignancy ratings

When comparing the segmented tumor masks with the ground truth tumor masks, it is evident that the model is relatively conservative when segmenting the border of the nodule and tends to have a rounder and more compact size. Six nodules in the test set received blank segmentation outputs, indicating that the model failed to delineate any tumor pixels. Four of these instances had a malignancy rating of 2, one had a rating of 1, and one had a rating of 4. Visual inspection confirmed that all six nodules were extremely small in the ground truth masks, suggesting that limited pixel area didn't allow for reliable feature extraction and led the model to suppress them as noise. The ground truth masks for these 6 instances are displayed in Figure 5.2, where row (A) is the malignancy rating and row (B) is the binary ground truth tumor mask.

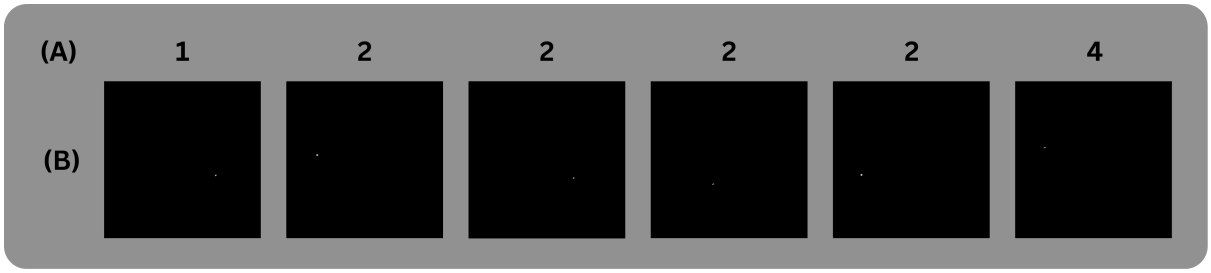

Figure 5.2: Ground truth masks for nodules with failed segmentation predictions

Performance Benchmarking

Compared to existing segmentation models reported in literature, the proposed model demonstrates competitive and superior performance. As presented in Table 2.1, DSC scores from recent studies range between 0.8074 and 0.8670, depending on the model architecture and input resolution. Despite operating with a higher input resolution of 512×512, the current model achieved a DSC of 0.8846 on the test set, exceeding that of UNet++ (0.8670) and other architectures used in the literature review. This result reflects the model's strong segmentation capabilities in the context of pulmonary nodule detection and shows that it performs at a level equal to or better than many state-of-the-art approaches in recent years.

Given the strong segmentation performance, the model serves as a reliable component within the full nodule detection pipeline. The accurately predicted binary masks provide consistent input to two branches of the MBNN architecture: the CNN-based mask image branch and the MLP-based semantic feature branch, where shape-based features are extracted directly from the predicted masks. The downstream impact of using these segmentation-derived masks, rather than ground truth annotations, is further evaluated in Section 5.2, where changes in predictive accuracy and error patterns are assessed.

Evaluation of Mask Drift Across Malignancy Ratings

To further evaluate the segmentation model beyond pixel-wise metrics, this section investigates mask drift by comparing semantic feature values extracted from the ground truth masks and the predicted masks. Specifically, it examines how the four shape-based semantic features (compactness, solidity, diameter, and spiculation proxy) change on average in the test set due to the segmentation model's output. These differences are analysed across the full test set and then separately by malignancy rating (1, 2, 4, and 5). This helps assess if the drift in feature values depends on malignancy severity and whether the model performs worse on tumors with certain malignancy levels.

The statistics reported represent the mean feature values calculated over the entire test set, pooled across all malignancy ratings. The segmented masks showed slightly higher compactness (0.9401 vs. 0.8859) and solidity (0.9231 vs. 0.9022), reflecting a tendency toward smoother and more regular contours. In contrast, the predicted spiculation proxy was notably lower (14.19 vs. 15.38), indicating less edge irregularity than seen in the annotated ground truth masks. While the average diameter remained relatively stable (19.19 vs. 19.31), this subtle underestimation suggests a mild contraction effect in the predicted mask boundaries. Together, these differences highlight a consistent structural simplification in the segmentation model's outputs.

Comparing feature values per malignancy rating reveals consistent trends in the segmentation model's behaviour. Across all categories, the predicted masks exhibit higher compactness and

lower spiculation proxy, suggesting a systematic smoothing effect and reduced edge irregularity regardless of malignancy severity. This simplification is most prominent in malignancy 1 and 2, where compactness is significantly overestimated and diameter is notably underestimated, especially for malignancy 1 (22.31 vs. 7.76). These cases likely reflect the model's difficulty in segmenting very small or ambiguous nodules, leading to collapsed or partial masks. For malignancy 4 and 5, the differences in diameter and spiculation proxy are less severe, indicating more reliable segmentation for larger or more clearly defined nodules. Overall, Table 7.3 - 7.7 in the Appendix (which contain the metrics refered to above) highlights that drift in shape-based features is not uniform and appears more pronounced in lower malignancy ratings.

5.2 Classification Model Performance

This section evaluates the classification performance of the baseline CNN and the MBNN using standard diagnostic metrics, including the confusion matrix, precision, recall, F1 score, and the ROC curve. The objective is to assess not only overall predictive accuracy but also how each model performs across different clinical scenarios. This analysis is particularly relevant for understanding where the MBNN architecture contributes to performance improvements with direct implications for diagnostic reliability.

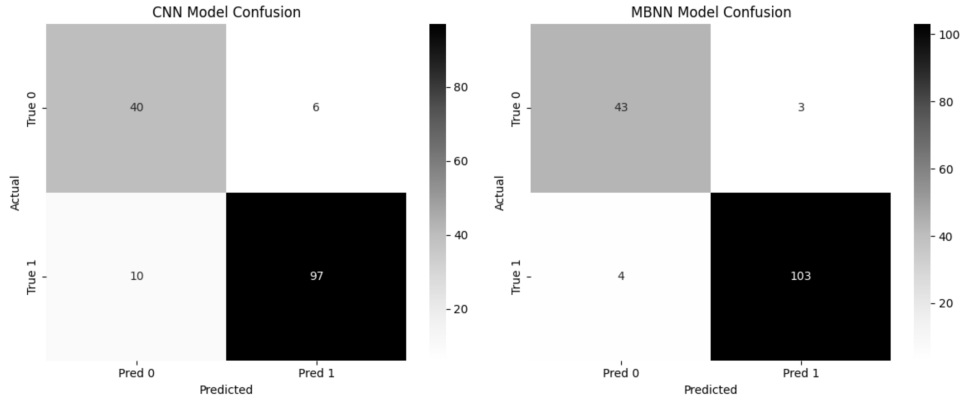

Figure 5.3: Confusion matrices for the baseline CNN and the multi-branch classification model

The confusion matrices for both the baseline CNN and the MBNN are shown in Figure 5.3, offering a clear picture of each model's strengths and weaknesses in terms of prediction outcomes. The baseline CNN misclassified 10 positive cases as negative, while the MBNN brought that number down to just 4. This represents a 60% reduction in false negatives, which is particularly significant in a clinical setting where missing a malignant case can delay critical treatment and lead to severe consequences, potentially leading to missed interventions at a crucial stage. The MBNN also reduced the number of false positives from 6 to 3, which amounts to a 50% decrease.

Malignancy Rating	**1**	**2**	**4**	**5**
Number in test set	21	25	48	59
Misclassified - CNN	2	2	7	3
Misclassified - MBNN	1	2	4	0

Table 5.1: Misclassification counts by malignancy rating (CNN vs MBNN)

Table 5.1 allows a more detailed inspection of each model's misclassification patterns across malignancy ratings. Notably, nodules with a malignancy rating of 4, remain the most error-prone category. The baseline CNN incorrectly classified 7 out of 48 of these cases (approximately 15%), while the MBNN reduced this to 4 (approximately 8%), representing a 57% reduction. For malignancy rating 5, which corresponds to the most clinically concerning cases, the CNN misclassified 3 of 59 nodules (5%), whereas the MBNN achieved 100% accuracy in this group.

These results indicate that the performance gains introduced by the MBNN are not only reflected in global metrics but are particularly concentrated in the more clinically critical categories. A reduction in false negatives at malignancy ratings 5 directly translates into improved diagnostic safety. Although the drop in false-positive predictions for the lower-risk groups (ratings 1 and 2) was modest (the misclassified instances in this group only decreased by 1), the clinical impact remains limited because any nodule flagged as malignant would undergo confirmatory biopsy before treatment decisions are made. The MBNN thus improves classification performance in a clinically meaningful way, addressing both sensitivity and specificity where it is most important.

Although false positives may be slightly less critical than false negatives in this context, they still carry significant emotional burden. Informing a patient that they might have cancer, only for a biopsy to later confirm it as benign, can cause unnecessary psychological distress and anxiety. These results demonstrate that the MBNN not only improves the model's predictive accuracy but also aligns more closely with the real-world demands of medical diagnosis, where both types of errors carry serious implications. The fact that such reductions in misclassifications are achieved by a model that builds on the original CNN structure, with only a moderate increase in number of trainable parameters, highlights the practical value of the MBNN design in improving diagnostic reliability.

		precision	recall	f1-score
CNN	0	0.8000	0.8696	0.8333
	1	0.9417	0.9065	0.9238
	accuracy			0.8954
	macro avg	0.8709	0.8881	0.8786
	weighted avg	0.8991	0.8954	0.8966
MBNN	0	0.9149	0.9348	0.9247
	1	0.9717	0.9626	0.9671
	accuracy			0.9542
	macro avg	0.9433	0.9487	0.9459
	weighted avg	0.9546	0.9542	0.9544

Table 5.2: CNN and MBNN Classification Report

The classification reports for both models, shown in Table 5.2, provide a more detailed perspective on performance by summarizing the precision, recall, and F1-score for each class. Looking first at the baseline CNN, there is a notable imbalance in the model's behavior between the two classes. For class 0 (benign), the CNN achieved a precision of 0.8000 and a recall of 0.8696, resulting in an F1-score of 0.8333. In contrast, for class 1 (malignant), the precision is substantially higher at 0.9417, with a recall of 0.9065 and an F1-score of 0.9238. These values confirm what the confusion matrix already suggested, that the CNN performs better at identifying positive (malignant) cases than negative ones, though it still leaves a concerning number of false negatives. When evaluating the macro and weighted averages, the F1-scores of 0.8786 and 0.8966 reflect solid overall performance but reveal that the model is not fully balanced across classes.

In comparison, the MBNN shows substantial improvements across all metrics and both classes. For class 0, precision rises from 0.8000 to 0.9149, and recall increases from 0.8696 to 0.9348, pushing the F1-score to 0.9247. Class 1 also sees improvement, with precision improving to 0.9717, recall to 0.9626, and F1-score to 0.9671. This consistency across both classes indicates that the MBNN not only improves general predictive capability but also addresses the class imbalance seen in the baseline model. The overall accuracy climbs from 89.5 percent to 95.4 percent, which represents a clear step up in performance. More importantly, the macro and weighted average scores are tightly aligned, each sitting above 0.94, which suggests that the MBNN maintains strong and balanced predictive power regardless of class.

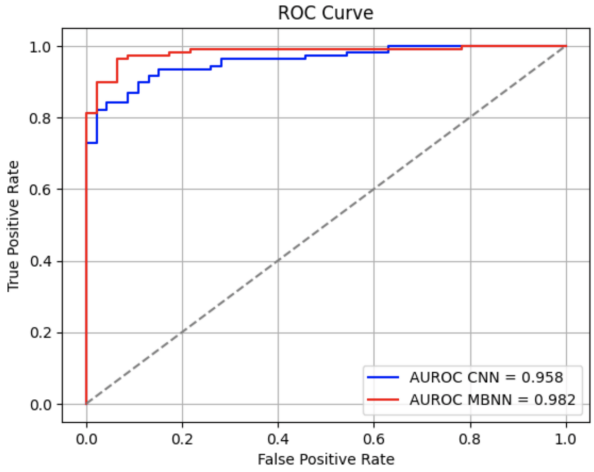

Figure 5.4: Receiver operating characteristic (ROC) curves for the baseline CNN and multi-branch classification model

The ROC curves for both models are presented in Figure 5.4, illustrating the relationship between the true positive rate and the false positive rate at varying threshold values. The curve associated with the MBNN shows an Area Under the ROC Curve (AUROC) of 0.982, while the baseline CNN's AUROC is slightly lower at 0.958. Both models show strong ability to differentiate between positive and negative cases, since scores closer to 1.0 mean better separation between the two classes. However, the MBNN's ROC curve is consistently closer to the upper-left corner of the plot, reflecting a more favorable balance between sensitivity and specificity across a wider range of thresholds. This subtle yet consistent improvement in the AUROC aligns with the improvement already observed in the confusion matrix and classification report. The MBNN not only performs better at a fixed threshold, but also offers more reliable behavior across varying thresholds, which can be particularly valuable in clinical scenarios where decision thresholds might be adjusted to prioritize sensitivity or specificity depending on the patient's context. A higher AUROC implies that the model is more robust to threshold selection and can still make confident distinctions. Together with the previous analyses, the ROC curve further supports the conclusion that the MBNN is a more reliable and clinically suitable model for lung cancer classification.

Classification performance utilizing segmentation output

To evaluate the impact of using segmentation-derived masks instead of ground truth masks within the MBNN framework, classification performance was compared across both input types.

		precision	recall	f1-score
Ground Truth	0	0.9149	0.9348	0.9247
	1	0.9717	0.9626	0.9671
	accuracy			0.9542
	macro avg	0.9433	0.9487	0.9459
	weighted avg	0.9546	0.9542	0.9544
Segmentation	0	0.8947	0.8293	0.8608
	1	0.9358	0.9623	0.9488
	accuracy			0.9252
	macro avg	0.9153	0.8958	0.9048
	weighted avg	0.9243	0.9252	0.9243

Table 5.3: Classification performance of the MBNN using ground truth and segmentation-derived masks

As shown in the classification reports in Table 5.3, overall performance decreased slightly when using the predicted masks. Accuracy dropped from 0.9542 to 0.9252, while macro-averaged F1-score declined from 0.9459 to 0.9048. The decrease was primarily driven by lower recall in class 0 (benign), which fell from 0.9348 to 0.8293, suggesting that the model missed more benign cases when masks were generated by the segmentation model. In contrast, performance for class 1 (malignant) remained stable, with only a minor drop in F1-score from 0.9671 to 0.9488. These results confirm that although classification performance slightly degrades when predicted masks are used, the model still maintains strong discriminatory capability, particularly for malignant nodules.

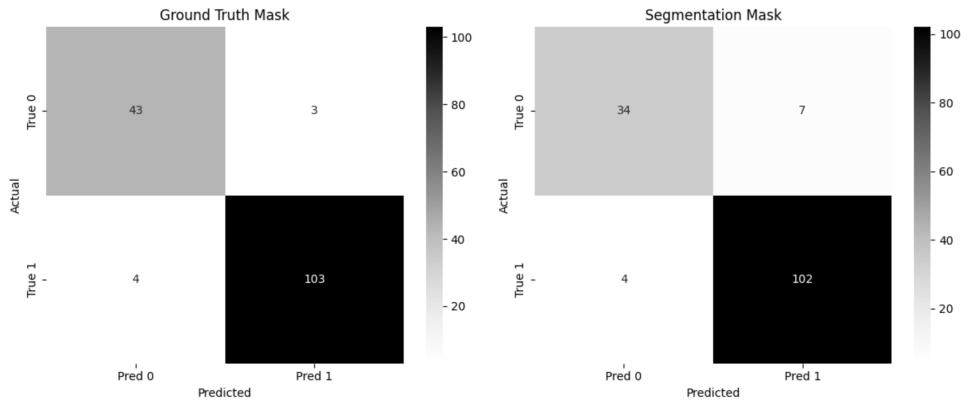

Figure 5.5: Confusion Matrices for MBNN Classification Using Ground Truth vs. Segmentation Masks

The confusion matrices in Figure 5.5 provide further detail on these shifts in performance. When ground-truth masks are used, the model records 43 true negatives, 103 true positives, 3 false positives, and 4 false negatives. Replacing the masks with segmentation outputs reduces the

true-negative count to 34 and increases false positives to 7, while the number of false negatives remains unchanged at 4 and true positives drop marginally to 102. The rise in false positives aligns with the decline in precision for class 0 noted in Table 5.3, indicating that benign nodules are more frequently misclassified as malignant when the mask quality is imperfect. Importantly, the stable false-negative count shows that sensitivity to malignant cases is largely preserved, confirming the minimal change observed in class 1 F1-score. Overall, the confusion-matrix patterns reinforce that the impact of mask drift is a modest increase in benign-to-malignant misclassification, whereas detection of malignant nodules remains robust.

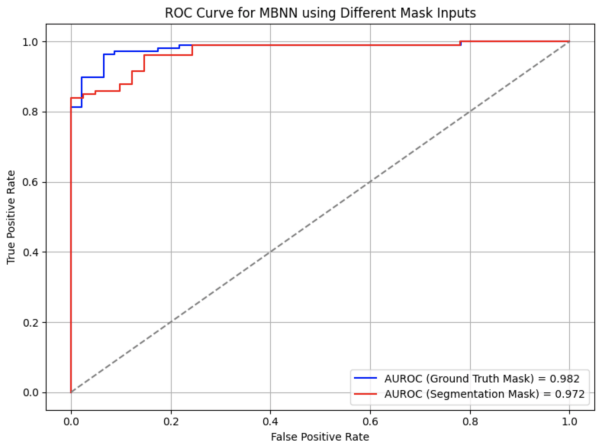

Figure 5.6: ROC Curve Comparison Using Ground Truth vs. Segmentation Masks as Input to MBNN

Continuing from the previous analysis, the ROC curve shown in Figure 5.6 provides a visual comparison of the classification performance when using ground truth masks versus segmentation model outputs as inputs into the MBNN. Both curves indicate strong discriminative capability, with AUROC values of 0.982 for the ground truth masks and 0.972 for the predicted masks.

The small gap between the two curves confirms that while the ground truth masks offer slightly better performance, the decrease introduced by using predicted masks is minimal. This aligns with earlier findings from the classification report, where performance was largely maintained despite a moderate drop in benign recall. The consistently high AUROC also reinforces the reliability of the segmentation model outputs as valid input for the classification pipeline, particularly in applications prioritizing malignant case detection.

Performance Benchmarking

To benchmark the performance of the proposed MBNN, results were compared against four models from the literature review, each using the LIDC-IDRI dataset. This benchmarking was done for the MBNN when using both radiologist-provided ground-truth masks and segmentation model–predicted masks as input. Two of the reference models use single-branch architectures,

while the other two are multi-branch networks.

The first benchmark model by Causey et al. (2018) tested two binary classification setups, both excluding nodules with a malignancy rating of 3. Their highest reported accuracy was 0.9620 for nodules rated 1 and 5, while grouping 1-2 as benign and 4-5 as malignant dropped accuracy to 0.9370. Our MBNN was trained on the malignancy range of (1-2 & 4-5) achieved 0.9542 accuracy, with macro and weighted F1 scores of 0.9459 and 0.9544, respectively. When using predicted segmentation masks, performance remained strong with 0.9252 accuracy and an AUROC of 0.972.

The second benchmark by Xu et al. (2020) used a 3D CNN trained on nodules rated 1 and 5, achieving an accuracy of 0.9445 and an AUROC of 0.976. Yuan et al. (2023) proposed a multi-branch 3D architecture combining image data and semantic features, reporting 0.9489 accuracy and an AUROC of 0.9784. Our MBNN, using ground-truth masks, outperformed both benchmarks with an accuracy of 0.9542 and an AUROC of 0.982. When using predicted segmentation masks instead, the accuracy dropped to 0.9252, which is notably lower than both Xu et al. and Yuan et al., but the AUROC remained comparable at 0.972. This suggests that although segmentation quality impacts classification accuracy, the model maintains a similar level of discriminatory power.

Zheng et al. (2024) proposed a multi-branch CNN that integrated grayscale images, segmentation masks, and texture maps, reaching 0.9011 accuracy and 0.9566 AUROC. Our MBNN achieved better results in both metrics and further included visual interpretability methods such as Grad-CAM, which their model lacked. Notably, even with predicted segmentation masks, our model outperformed theirs.

Overall, these comparisons demonstrate that the proposed MBNN not only matches but exceeds the accuracy of volumetric CNNs and other MBNN approaches, but also delivers greater interpretability and robustness across metrics.

5.3 Case-Based Analysis of Classification Outcomes

Understanding where and why deep learning models succeed or fail is key to improving performance and building trust in their predictions. This section analyzes the classification outcomes of both the baseline CNN and the MBNN, with a focus on cases where the models agree or disagree with the ground truth. By categorizing test set results into four outcome types, the analysis provides insight into each model's decision-making behaviour.

- Case 1: CNN Correct, MBNN Incorrect
- Case 2: CNN Incorrect, MBNN Correct
- Case 3: Both Models Incorrect
- Case 4: Both Models Correct

From the set of instances belonging to each case, a single representative instance is selected using the Mahalanobis Distance calculated on the semantic feature values. This ensures that the instance chosen to investigate is centrally located within the distribution of the case subset, and therefore represents the typical characteristics of instances in that specific case. The selected instance is then examined using the binary tumor mask, CT scan image, and Grad-CAM activation maps. This analysis aims to uncover visual or feature-based factors that can explain the difference in model performance.

In addition to the qualitative analysis, a statistical analysis of the semantic features is conducted for the 2 extreme cases, when both models are correct and when both are incorrect. For each semantic feature, the mean, standard deviation, minimum, maximum, and median values are calculated. This analysis aims to determine whether certain tumor shape characteristics are associated with higher or lower classification accuracy.

By utilizing both quantitative and qualitative assessment, the case-based analysis aims to reveal underlying causes of disagreement between model predictions and their classification errors.

Case 1: CNN Correct, MBNN Incorrect

This is the only instance in the test set where the CNN correctly classifies the tumor while the MBNN does not. The corresponding CT image displays a small and clearly distinguishable region located centrally. The binary tumor mask confirms the regions boundaries and indicates that the segmentation is clean and well-defined, with no apparent artifacts or irregularities.

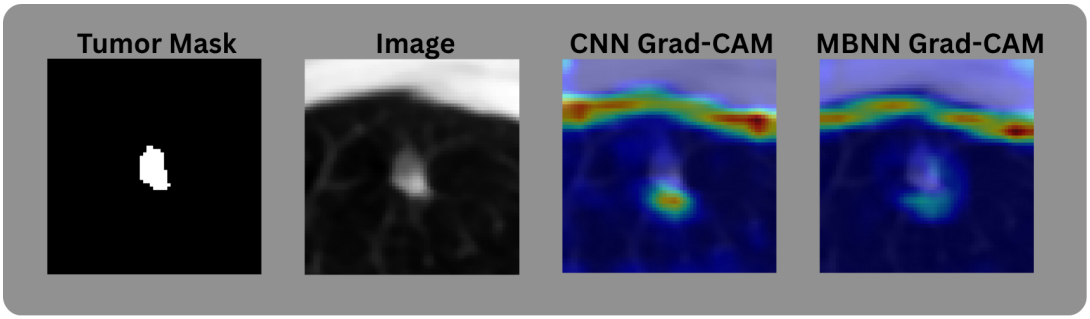

Figure 5.7: Grad-CAM visualisation for Case 1: CNN correct, MBNN incorrect

The Grad-CAM visualization for the CNN model reveals two primary regions of attention: the tumor itself and a non-tumor region in the upper part of the image, likely near the pleura.

The models focus on the tumor region indicates that is it using relevant local features to inform its decision, but the additional attention at the top of the image suggests that the model is being influenced by contextual or structural elements outside of the tumor region. This divided attention contributes to the relatively low classification confidence of 57.65%.

In contrast, the Grad-CAM output of the MBNN shows minimal activation over the tumor area. The model predominantly focuses on the non-tumor region in the upper part of the image, with little emphasis on the actual lesion. This suggests that that the contextual features outside of the tumor region led the MBNN to incorrectly classify the tumor as Malignant. Despite incorporating the baseline CNN as one of its branches, the MB model failed to reproduce the correct prediction, which suggests that the feature fusion process may have diluted the relevant information from the CNN branch and either the semantic feature MLP branch or the tumor mask CNN pushed the prediction into the wrong direction.

This case represents the only instance where the baseline CNN outperformed the MBNN. While the CNN correctly classified the tumor as benign, it did so with very low confidence, suggesting uncertainty in its prediction. The MBNN, despite using additional inputs, misclassified the instance with higher confidence. This outcome highlights a rare but important limitation of the MBNN architecture, where the integration of multiple modalities may suppress useful signals and amplify misleading contextual features, in some cases.

Case 2: CNN Incorrect, MBNN Correct

This instance is one of 10 in the test set where the MBNN correctly classifies the tumor while the CNN does not. The corresponding CT image displays a clearly distinguishable region located centrally. The binary mask confirms accurate segmentation, clearly outlining the lesion with no visible artifacts or segmentation errors.

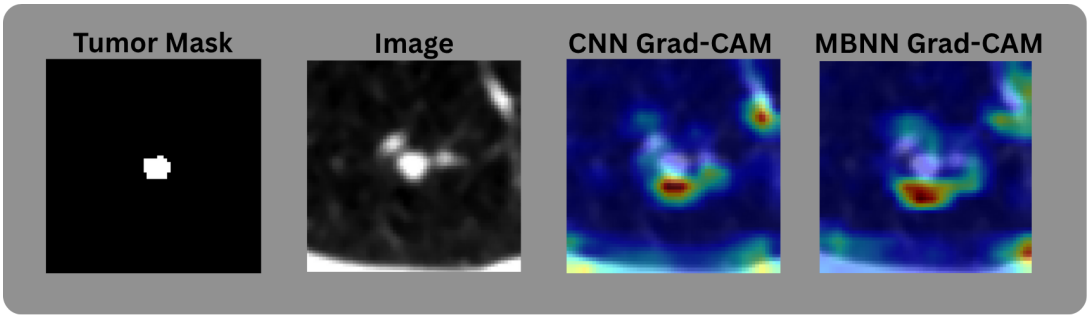

Figure 5.8: Grad-CAM visualisation for Case 2: CNN incorrect, MBNN correct

The Grad-CAM visualization for the CNN model shows highly concentrated attention directly surrounding the bottom area of the tumor region. While attention near the periphery of the tumor can be informative for identifying malignant characteristics such as spiculated or invasive growth, in this case there is no visible evidence of branching or structural irregulari-

ties in that region. As a result, this intense focus is misplaced and does not reflect clinically relevant features. Additionally, there is a second region of attention visible near the top right where a distinct bright structure is present. While not being directly related to the tumor, this secondary focus can be interpreted as the model correctly identifying other high-intensity regions. Despite this, the CNN predicted the tumor as benign with high confidence (97.34%), which suggests that its interpretation of relevant features was ultimately inaccurate.

In contrast, the Grad-CAM output for the MBNN shows a broader and more context aware attention map. While it still includes some focus below the tumor, it more evenly highlights regions around the lesion, particularly the top right and top left, where branching structures are visible. This distribution suggests the model is interpreting the tumour environment rather than fixating on isolated points. The inclusion of the tumor mask and semantic features likely guides this behaviour, helping the model attend to clinically relevant regions and leading to a correct malignant classification.

The MBNN also shows a small red hotspot in the bottom right corner, indicating a flaw in its attention that could be addressed in future work and is discussed further in Chapter Six.

Case 3: Both Models Incorrect

This instance is one of six in the test set where both the CNN and the MBNN model misclassify the tumor. The CT image shows a clearly visible tumor with subtle but noticeable branching extending from the central mass. The corresponding tumor mask, which is the ground truth provided by radiologists, captures the core of the tumor but does not fully define the branching structures seen in the image. As a result, the mask provides an incomplete representation of the tumor's spatial context and limits the effectiveness of the MBNN relying on this input to improve on the baseline CNN.

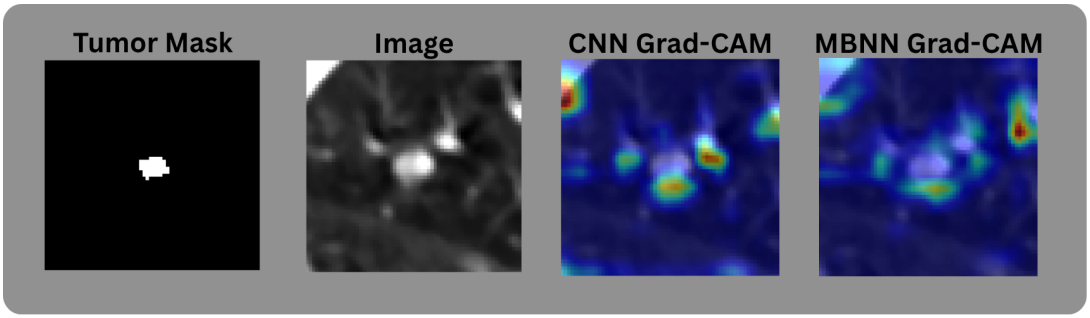

Figure 5.9: Grad-CAM visualisation for Case 3: inCNN correct, MBNN incorrect

The Grad-CAM overlays for this instance further illustrate key differences in how the CNN and MBNN models interpret the image. The MBNN model displays a broader and more spatially distributed attention pattern. Activation is spread around the entire tumor region, including peripheral areas that extend beyond the core, with a particularly strong focus on the right side.

This more comprehensive attention distribution suggests that the model is leveraging a wider contextual understanding when evaluating the tumor, likely supported by the additional input from the tumor mask and semantic features.

In contrast, the CNN model exhibits more narrowly concentrated attention with several isolated activation spots. A strong activation appears in the top left corner of the image, far from the tumor, suggesting that the model may be reacting to irrelevant visual features such as the pleura, which lies along the outer edge of the lung. The attention within the tumor region is more fragmented and less extensive than that of the MBNN. This focused but misaligned attention pattern likely contributed to the model's incorrect benign classification, despite its extremely high confidence (99.64%).

Together, these observations reinforce that while both models misclassified the tumor, the MBNN model demonstrated a more appropriate and distributed focus across the lesion. A notable shortcoming of the MBNN is that it remains limited by the quality of the tumor mask. In cases like this where the mask provides an incomplete representation of the tumor's spatial context, specifically with respect to peripheral branching, the model may still be led to an incorrect classification. Introducing an additional input branch capable of capturing broader contextual or structural information, such as texture-based spatial features, could help guide the model towards more accurate predictions in such scenarios, and is discussed further in Chapter 6. However, the MBNN clearly improves upon the baseline CNN by building on its outputs and enhancing them through feature fusion. Since the goal of this thesis is to explore whether incorporating tumor-specific features and spatial context improves classification, this outcome demonstrates the model's potential to do so ,while also identifying directions for future work.

Case 4: Both Models Correct

This instance is part of the majority case in the test set, where both the CNN and MBNN models correctly classify the tumor. Visual inspection of the Grad-CAM overlays reveals that both models focus on broadly similar regions, particularly focusing on the tumor and its immediate surroundings. This alignment in attention likely reflects shared underlying features learned during training, especially since the MBNN model incorporates the CNN pathway as one of its branches.

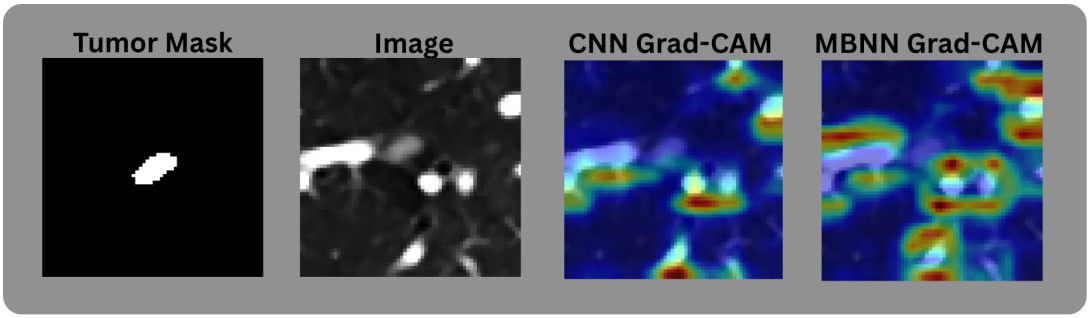

Figure 5.10: Grad-CAM visualisation for Case 4: CNN correct, MBNN correct

However, it is evident that the MBNN's attention is more holistically distributed around the lesion, covering a larger area that includes both the tumor core and its periphery. In contrast, the CNN's activation appears to be more fragmented and localized in certain regions, focusing on distinct spots. The MBNN's wider attention appears to reflect a stronger spatial awareness, capturing not only key features within the tumor but the surrounding regions as well, which are important for assessing malignancy.

This difference is clinically significant, as malignant tumors tend to exhibit irregular, infiltrative growth patterns that extend beyond the visible core of the lesion. The MBNN's broader attention improves its overall spatial awareness, placing it in a stronger position to detect subtle cues that may indicate malignancy. The integration of the tumor mask and semantic feature branches reinforces this enhanced awareness, helping the model interpret the lesion in the context of its surrounding structure rather than in isolation.

Statistical Analysis of Semantic Features

To investigate how tumour morphology influences classification performance, summary statistics and pair-wise correlation matrices were compared for the two extreme agreement cases, namely case 3 where both models misclassified and case 4 where both models classified correctly.

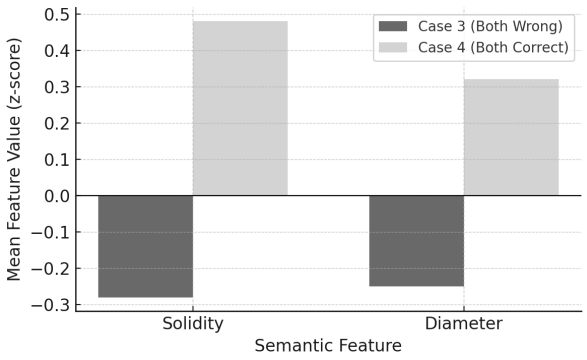

Figure 5.11: Mean z-scored semantic feature values for correctly and incorrectly classified tumors (Cases 3 and 4)

In terms of average feature values, tumours in Case 4 displayed characteristics more strongly associated with malignancy, particularly in terms of solidity and diameter. Specifically, the mean solidity increased from –0.2761 in Case 3 to +0.4821 in Case 4, and the mean diameter rose from –0.2469 to +0.3193. These differences are visualised in Figure 5.11, and suggest that larger and denser tumours are more readily captured by the classification model and are easier for the model to correctly classify. The complete tables containing all average semantic feature values for case 3 and 4 can be seen in Table 7.8 & 7.9 in the Appendix.

More informative differences emerged in the correlation structure between features.

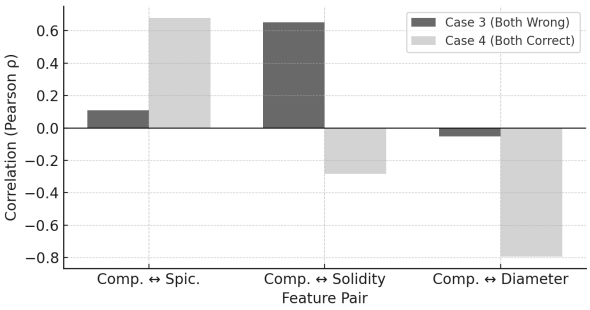

Figure 5.12: Pairwise semantic feature correlations for correctly and incorrectly classified tumors (Cases 3 and 4)

In Case 4, where classification was successful, compactness and spiculation proxy were strongly positively correlated (0.6788), compared to a near-zero correlation in Case 3 (0.1101). This relationship is shown in Figure 5.12, and is consistent with tumors that have thin, radial branches: as spiculation increases (more negative values in the proxy), compactness decreases due to the irregular margin. Compactness and solidity also show a reversal in correlation sign, changing from a strong positive correlation in Case 3 (0.6510) to a weak negative correlation in Case 4 (–0.2834). This indicates that in misclassified tumours, the mask captures only the dense core, resulting in features that describe a similar structure. When branches are included (Case 4), compactness drops due to contour irregularity, while solidity remains high, capturing different morphological aspects. Finally, compactness and diameter show a strong negative correlation in Case 4 (–0.7933), compared to no clear relationship in Case 3 (–0.0517), suggesting that correctly segmented large tumours tend to exhibit irregular perimeters consistent with branching. The complete tables containing all semantic feature correlation matrices for case 3 and 4 can be seen in Table 7.10 & 7.11 in the Appendix.

These results imply that the success of the multi-branch network depends not only on the presence of informative features but also on the diversity between them. In correctly classified cases, each semantic feature encodes complementary shape information, improving the effectiveness of feature fusion across branches. In contrast, when segmentation under-represents the tumour's peripheral complexity, the semantic features become redundant, and both the binary

mask and semantic feature branches deliver overlapping, less informative inputs. These findings highlight the importance of accurate segmentation in multimodal classification architectures and highlight the need for improved representation of tumour margins and branching.

Chapter 6

Conclusion and Future Work

This chapter concludes the thesis by summarising the main contributions, presenting key findings in relation to the research questions, and outlining limitations and directions for future work. The aim is to reflect on what was achieved and identify areas that could benefit from further development, particularly in the context of real-world deployment and clinical use.

6.1 Summary of Contributions

This thesis presents a fully automated, interpretable pipeline for pulmonary nodule classification using CT scans.
Key contributions include:

- Development of an end-to-end system combining segmentation, shape-based feature extraction, classification, and visual explanation.
- Introduction of a multi-branch classification network that integrates image, mask, and semantic features for improved performance and interpretability.
- Evaluation of how segmentation output quality affects downstream prediction and feature behaviour, highlighting trade-offs in automation.

6.2 Key Findings

This section first presents the main findings in direct response to the five research questions. Additional key results that fall outside the scope of these questions are then discussed separately below.
Research question findings:

1. The multi-branch neural network (MBNN) outperforms the baseline CNN across all classification metrics. Accuracy improves from 89.5% to 95.4%, with AUROC increasing from 0.958 to 0.982. The MBNN significantly reduces false negatives from 10 to 4 and false

positives from 6 to 3, showing that it is both more sensitive and specific. These improvements come with only a modest increase in the number of trainable parameters (from 2.19 million to 2.25 million), suggesting the added branches, specifically those incorporating mask and semantic features, contribute meaningful information without adding unnecessary complexity.

2. Grad-CAM heatmaps reveal that the baseline CNN often focuses on both the nodule and unrelated surrounding tissue, such as bright areas near the pleura, which likely contributes to its misclassifications. In contrast, the MBNN consistently concentrates attention within the tumor boundary and shows a more holistic focus across the entire lesion. It pays greater attention to peripheral branching patterns, an important visual cue for malignancy, suggesting that the mask and semantic branches help guide the model toward more clinically meaningful features. This improved focus not only enhances classification accuracy but also makes the model's decisions more interpretable and aligned with how radiologists assess nodules.

3. Semantic feature analysis reveals that correctly classified tumours (Case 4) have higher average values for features like solidity and diameter, which are more commonly associated with malignancy. However, the more informative insight comes from the correlation patterns between features. In Case 4, strong correlations such as between compactness and spiculation proxy or compactness and diameter suggest that the features capture distinct and complementary shape characteristics. In contrast, in misclassified cases (Case 3), these correlations are weak or absent, indicating that the features provide overlapping or less informative inputs. This suggests that classification performance improves when semantic features have diverse and non-redundant morphological signals.

4. The U-Net++ segmentation model achieved a Dice score of 0.8846 and an IoU of 0.8038, exceeding the performance of benchmarked segmentation models, including UNet++ variants and others reporting scores in the 0.8074–0.8670 range. The perfromacne metrics of the MBNN, discussed in research question one, outperformed all non-transformer classification models in the literature. Replacing radiologist-provided ground-truth masks with the segmentation model's predicted masks leads to a slight drop in performance, with accuracy falling from 95.4% to 92.5% and AUROC from 0.982 to 0.958. Most of the performance loss comes from reduced recall on benign cases, which drops from 93.5% to 82.9%. Importantly, the number of false negatives for malignant cases remains unchanged, meaning the model's ability to detect malignancies is preserved. These results suggest that the fully automated pipeline remains clinically feasible, with only a modest and acceptable trade-off in performance when manual masks are removed from the process.

5. When using predicted masks, the semantic features tend to show slightly higher compactness and solidity and a lower spiculation proxy across all malignancy ratings, indicating smoother and more simplified contours compared to ground-truth masks. This effect is

most pronounced in malignancy ratings 1 and 2, where compactness is notably overestimated and diameter is significantly underestimated, especially in malignancy 1, where the average diameter drops from 22.31 to 7.76. These differences suggest the segmentation model struggles with very small or ambiguous nodules, often producing partial or collapsed masks. In higher malignancy ratings (4 and 5), the differences in semantic features are smaller, indicating more reliable segmentation for larger or more distinct tumours. This suggests that even small shifts in mask shape can influence the informativeness of the extracted features and, affect classification reliability.

Additional findings:

- **Stable detection of malignant nodules with predicted masks rather than ground truth masks.** Recall for malignant cases remained high ($0.9626 \rightarrow 0.9488$) and the number of false negatives was unchanged, indicating that the clinically critical task of finding cancerous nodules is robust to segmentation drift.
- **Rise in benign false positives with predicted masks.** The principal cost of using automatic masks was an increase in benign-to-malignant misclassifications (false positives rose from 3 to 7), reflected in a precision drop for class 0. This trade-off favours patient safety by avoiding missed cancers while imposing only a moderate burden of follow-up scans.

6.3 Limitations

The following limitations were identified during the study:

- **Dependence on segmentation quality.** The classification model's performance is directly influenced by the quality of the segmentation output. Incorrectly segmented masks, can distort the semantic feature values and reduce classification accuracy.
- **Incomplete radiologist-annotated masks may omit critical structure.** In cases such as Case 3, the ground truth mask annotated by the radiologist excluded peripheral branches of the tumour, retaining only the core mass. This incomplete representation resulted in the MBNN misclassifying the nodules, highlighting how even expert-drawn masks can limit downstream model performance when key morphological details are omitted.
- **Reduced generalisation on ambiguous or small nodules.** The model occasionally failed to segment or classify very small nodules with low pixel area, likely due to weak feature signals and low resolution of critical patterns.

6.4 Future Research

- **Incorporate a branching descriptor into the feature set:** Future work should explore shape descriptors that explicitly quantify tumour branching, such as the number,

length, or irregularity of offshoots. These metrics could improve classification performance in edge cases where traditional shape features like compactness and solidity fail to capture fine-grained morphology.

- **Explore joint training of segmentation and classification:** End-to-end optimisation, where segmentation quality is directly linked to classification loss, could reduce error propagation and encourage masks that preserve clinically relevant detail.
- **Create Confidence score for segmentation:** Use the confidence score of the segmentation model to determine the weighting of each of the three branches in the MBNN. If the segmentation is not confident then the CT image branch should receive the highest weighting as it is not affected by the biased segmentation model output.
- **Prioritise sensitivity in clinical settings:** Future work could explore training strategies that explicitly minimise false negatives, even at the expense of increased false positives. By tuning the model to favour recall over precision, especially in ambiguous cases, it may be possible to capture all malignant nodules. When combined with expert radiological review and follow-up procedures such as biopsy, this approach could help ensure that no cancerous cases are missed in practice.

Taken together, these findings offer a practical step toward interpretable, fully automated nodule classification systems with reduced reliance on expert annotations. Beyond answering the initial research questions, this work opens a clear path for future research, particularly around better capturing morphological complexity and improving segmentation-classification performance. With continued development, the proposed framework can form the backbone of a highly robust and trustworthy pipeline suitable for real clinical deployment.

Appendix

Level	**Layer(s)**	**Number of Filters**	**Output Size**
Encoder Block 1	conv0_0	64	512 × 512
Encoder Block 2	conv1_0	128	256 × 256
Encoder Block 3	conv2_0	256	128 × 128
Encoder Block 4	conv3_0	512	64 × 64
Encoder Block 5	conv4_0	1024	32 × 32
Decoder Pathway	conv0_1 to conv0_4	64	512 × 512
	conv1_1 to conv1_3	128	256 × 256
	conv2_1 to conv2_2	256	128 × 128
	conv3_1	512	64 × 64
Final Output Layer	1×1 Conv (OutConv)	1	512 × 512

Table 7.1: U-Net++ Model Architecture with Filter Counts and Output Sizes

Branch / Level	**Layer(s)**	**Filters / Units**	**Output Size**
Image Branch			
Conv 1	Conv2D 3×3 + ReLU	32	64×64
Pool 1	MaxPool 2×2	–	32×32
Conv 2	Conv2D 3×3 + ReLU	64	32×32
Pool 2	MaxPool 2×2	–	16×16
Conv 3	Conv2D 3×3 + ReLU	128	16×16
Pool 3	MaxPool 2×2	–	8×8
Flatten	—	–	8192
Dense	Dense + ReLU + Dropout	256	256
Mask Branch			
Conv 1	Conv2D 3×3 + ReLU	16	64×64
Pool 1	MaxPool 2×2	–	32×32
Conv 2	Conv2D 3×3 + ReLU	32	32×32
Pool 2	MaxPool 2×2	–	16×16
Conv 3	Conv2D 3×3 + ReLU	64	16×16
Pool 3	MaxPool 2×2	–	8×8
Flatten	—	–	4096
Dense	Dense + ReLU + Dropout	64	64
Semantic Feature Branch			
Input	4-D semantic vector	4	4
Dense 1	Dense + ReLU + Dropout	32	32
Dense 2	Dense + ReLU + Dropout	16	16
Fusion and Output			
Concatenate	Image 256 ‖ Mask 64 ‖ Sem. 16	336	336
Dense 1	Dense + ReLU + Dropout	64	64
Output	Dense + Sigmoid	1	1

Table 7.2: Layer-wise architecture and output dimensions of the multi-branch neural network (MBNN)

Semantic Feature	**Malignancy Rating 1**	
	Segmentation Mask	**Ground-Truth Mask**
spiculation_proxy	10.7110	15.7962
compactness	1.2173	0.8397
solidity	0.9445	0.9100
diameter	7.7578	22.3097

Table 7.3: Mean semantic-feature values for malignancy rating 1: comparison of segmentation-derived vs. ground-truth masks.

Semantic Feature	**Malignancy Rating 2**	
	Segmentation Mask	**Ground-Truth Mask**
spiculation_proxy	14.6534	16.1534
compactness	0.9120	0.8384
solidity	0.9147	0.8873
diameter	17.8193	18.2454

Table 7.4: Mean semantic-feature values for malignancy rating 2: comparison of segmentation-derived vs. ground-truth masks.

Semantic Feature	**Malignancy Rating 4**	
	Segmentation Mask	**Ground-Truth Mask**
spiculation_proxy	14.0846	15.0788
compactness	0.9518	0.9100
solidity	0.9289	0.9073
diameter	17.5164	17.5383

Table 7.5: Mean semantic-feature values for malignancy rating 4: comparison of segmentation-derived vs. ground-truth masks.

Semantic Feature	Malignancy Rating 5	
	Segmentation Mask	**Ground-Truth Mask**
spiculation_proxy	14.1213	15.2081
compactness	0.9451	0.8992
solidity	0.9132	0.9008
diameter	20.1338	20.0902

Table 7.6: Mean semantic-feature values for malignancy rating 5: comparison of segmentation-derived vs. ground-truth masks.

Semantic Feature	All Malignancy Ratings	
	Segmentation Mask	**Ground-Truth Mask**
spiculation_proxy	14.1851	15.3818
compactness	0.9401	0.8859
solidity	0.9231	0.9022
diameter	19.1932	19.3127

Table 7.7: Mean semantic-feature values across all malignant nodules: comparison of segmentation-derived vs. ground-truth masks.

Semantic Features	mean	std	min	max	median
spiculation_proxy	-0.2980	0.1048	-0.3530	-0.0852	-0.3365
compactness	-0.2330	0.1882	-0.4750	0.0445	-0.1996
solidity	-0.2761	0.5527	-0.9318	0.7016	-0.4077
diameter	-0.2469	0.5336	-1.3167	0.0883	-0.0737

Table 7.8: Semantic Feature Statistics Case 3

Semantic Features	mean	std	min	max	median
spiculation_proxy	-0.3533	0.0734	-0.5315	-0.1643	-0.3717
compactness	-0.2264	0.6264	-3.3194	0.5962	-0.1051
solidity	0.4821	1.0324	-0.9712	2.8854	0.3154
diameter	0.3193	0.9264	-1.0728	5.6507	0.2980

Table 7.9: Semantic Feature Statistics Case 4

	spiculation_proxy	compactness	solidity	diameter
spiculation_proxy	1.0000	0.1101	-0.6367	-0.9961
compactness	0.1101	1.0000	0.6510	-0.0517
solidity	-0.6367	0.6510	1.0000	0.6819
diameter	-0.9961	-0.0517	0.6819	1.0000

Table 7.10: Semantic Feature Correlation Matrix Case 3

	spiculation_proxy	compactness	solidity	diameter
spiculation_proxy	1.0000	0.6788	-0.7780	-0.9066
compactness	0.6788	1.0000	-0.2834	-0.7933
solidity	-0.7780	-0.2834	1.0000	0.6955
diameter	-0.9066	-0.7933	0.6955	1.0000

Table 7.11: Semantic Feature Correlation Matrix Case 4

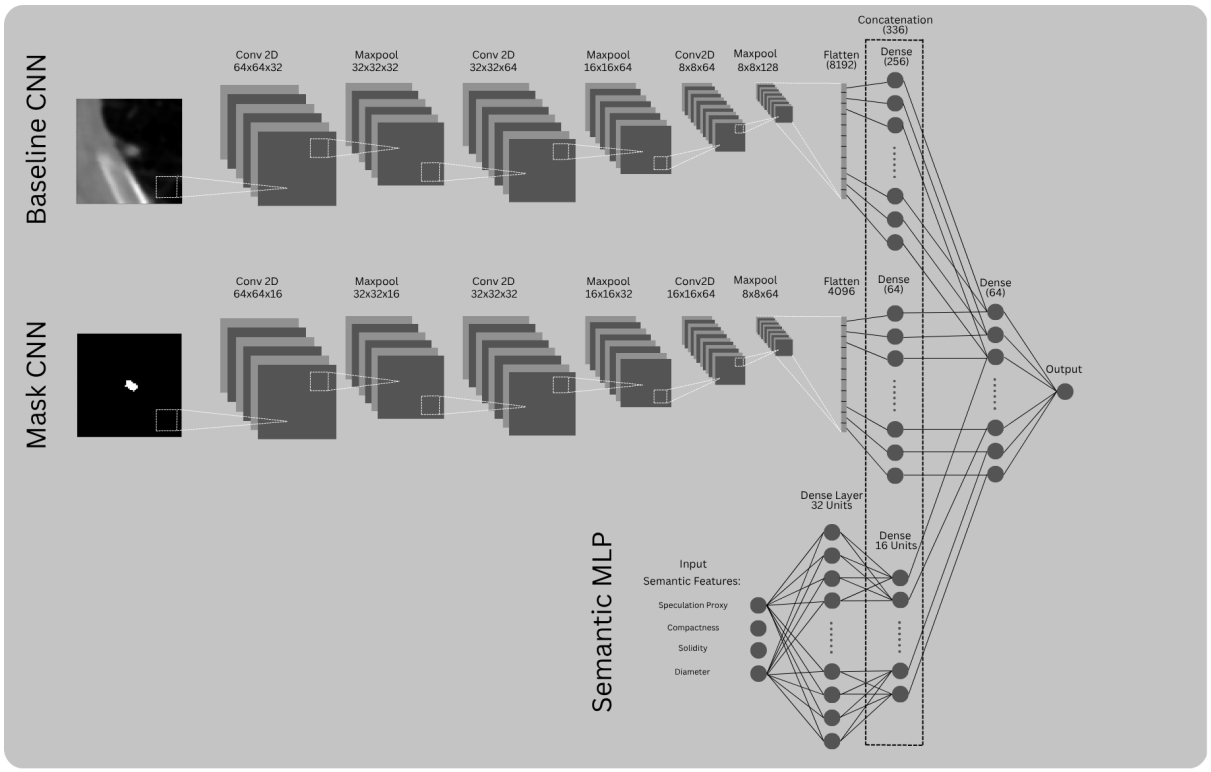

Figure 7.1: MBNN complete design

Bibliography

Abid, M. M. N., Zia, T., Ghafoor, M., & Windridge, D. (2021). Multi-view convolutional recurrent neural networks for lung cancer nodule identification. *Neurocomputing*, *453*, 299–311.

Aerts, H. J., Wee, L., Rios Velazquez, E., Leijenaar, R. T., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., et al. (2019). Data from nsclc-radiomics. *(No Title)*.

Ali, I., Muzammil, M., Haq, I. U., Amir, M., & Abdullah, S. (2020). Efficient lung nodule classification using transferable texture convolutional neural network. *Ieee Access*, *8*, 175859–175870.

American Cancer Society. (2024). Cancer facts figures 2024 [Provides five-year survival rates for lung cancer by stage].

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Shpanskaya, K., Corrado, G. S., Naidich, D. P., & Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, *25*, 954–961. https://doi.org/10.1038/s41591-019-0447-x

Armato III, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., Zhao, B., Aberle, D. R., Henschke, C. I., Hoffman, E. A., et al. (2011). The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans. *Medical physics*, *38*(2), 915–931.

Callister, M. E. J., Tonge, J., & Baldwin, D. R. (2024). Artificial intelligence and lung cancer screening: The state of the art. *Clinical Radiology*, *79*, 82–90. https://doi.org/10.1016/j.crad.2023.11.002

Causey, J. L., Zhang, J., Ma, S., Jiang, B., Qualls, J. A., Politte, D. G., Prior, F., Zhang, S., & Huang, X. (2018). Highly accurate model for prediction of lung nodule malignancy with ct scans. *Scientific reports*, *8*(1), 9286.

Duffy, S. W., Field, J. K., Baldwin, D. R., Hansell, D. M., Devaraj, A., & Wald, N. J. (2023). Use of artificial intelligence to support lung cancer screening. *Thorax*, *78*(1), 5–6. https://doi.org/10.1136/thorax-2022-219470

Ferlay, J., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, M., Znaor, A., & Soerjomataram, I. (2024). Global cancer statistics 2022: Globocan estimates of incidence and

mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *74*(2), 223–248. https://doi.org/10.3322/caac.21834

Kumar, S. N., Bruntha, P. M., Daniel, S. I., Kirubakar, J. A., Kiruba, R. E., Sam, S., & Pandian, S. I. A. (2021). Lung nodule segmentation using unet. *2021 7th International conference on advanced computing and communication systems (ICACCS)*, *1*, 420–424.

Lin, J., She, Q., & Chen, Y. (2023). Pulmonary nodule detection based on ir-unet++. *Medical & Biological Engineering & Computing*, *61*(2), 485–495.

Liu, C., Zhao, R., & Pang, M. (2023). Semantic characteristic grading of pulmonary nodules based on deep neural networks. *BMC Medical Imaging*, *23*(1), 156.

Nibali, A., He, Z., & Wollersheim, D. (2017). Pulmonary nodule classification with deep residual networks. *International journal of computer assisted radiology and surgery*, *12*, 1799–1808.

Ren, Y., Tsai, M.-Y., Chen, L., Wang, J., Li, S., Liu, Y., Jia, X., & Shen, C. (2020). A manifold learning regularization approach to enhance 3d ct image-based lung nodule classification. *International journal of computer assisted radiology and surgery*, *15*, 287–295.

Royal College of Radiologists. (2024). State of the wait: Diagnostic imaging [Reports on CT/MRI reporting delays and workforce shortages in the UK]. Policy Report.

Setio, A. A. A., Traverso, A., De Bel, T., Berens, M. S., Van Den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M. E., Geurts, B., et al. (2017). Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The luna16 challenge. *Medical image analysis*, *42*, 1–13.

Tang, H., Zhang, C., & Xie, X. (2019). Nodulenet: Decoupled false positive reduction for pulmonary nodule detection and segmentation. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, 266–274.

Tyagi, S., & Talbar, S. N. (2022). Cse-gan: A 3d conditional generative adversarial network with concurrent squeeze-and-excitation blocks for lung nodule segmentation. *Computers in Biology and Medicine*, *147*, 105781.

Wang, P., Ge, J., Zheng, D., Zhu, X., Liu, J., Wu, Y., Lu, L., Yan, S., Jin, D., & Ye, X. (2023). Anatomy-guided deep learning model for accurate and robust gross tumor volume segmentation in lung cancer radiation therapy. *International Journal of Radiation Oncology, Biology, Physics*, *117*(2), e71.

Wang, S., Zhou, M., Liu, Z., Liu, Z., Gu, D., Zang, Y., Dong, D., Gevaert, O., & Tian, J. (2017). Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. *Medical image analysis*, *40*, 172–183.

Wu, K., Peng, B., & Zhai, D. (2022). Multi-granularity dilated transformer for lung nodule classification via local focus scheme. *Applied Sciences*, *13*(1), 377.

Xu, X., Wang, C., Guo, J., Gan, Y., Wang, J., Bai, H., Zhang, L., Li, W., & Yi, Z. (2020). Mscs-deepln: Evaluating lung nodule malignancy using multi-scale cost-sensitive neural networks. *Medical Image Analysis*, *65*, 101772.

Yuan, H., Wu, Y., & Dai, M. (2023). Multi-modal feature fusion-based multi-branch classification network for pulmonary nodule malignancy suspiciousness diagnosis. *Journal of Digital Imaging*, *36*(2), 617–626.

Zhang, S., Sun, F., Wang, N., Zhang, C., Yu, Q., Zhang, M., Babyn, P., & Zhong, H. (2019). Computer-aided diagnosis (cad) of pulmonary nodule of thoracic ct image using transfer learning. *Journal of digital imaging*, *32*, 995–1007.

Zheng, R., Wen, H., Zhu, F., & Lan, W. (2024). Attention-guided deep neural network with a multichannel architecture for lung nodule classification. *Heliyon*, *10*(1).

Zhou, Z., Li, J., Wang, X., & Wang, X. (2021). Deep learning for lung cancer detection: Taxonomy, survey and future directions. *Pattern Recognition*, *119*, 108071. https://doi.org/10.1016/j.patcog.2021.108071