Department of Electrical and Computer Engineering
North South University

Senior Design Project

Cross-Architecture Knowledge Distillation Framework with Multi-scale Geometric Feature Fusion in Medical Image Classification and Segmentation

MD. Shafat Islam Khan **2121517642**

Mahiyat Nawar Mantaqa **2122455042**

Faculty Advisor:

Dr. Mohammad Shifat-E-Rabbi

Assistant Professor

ECE Department

Spring, 2025

LETTER OF TRANSMITTAL

April, 2025

To

Dr. Mohammad Abdul Matin
Chairman,
Department of Electrical and Computer Engineering
North South University, Dhaka

Subject: **Submission of Capstone Project Report on "Cross-Architecture Knowledge Distillation Framework with Multi-scale Geometric Feature Fusion in Medical Image Classification and Segmentation"**

Dear Sir,

With due respect, we would like to submit our **Capstone Project Report** on **"Cross-Architecture Knowledge Distillation Framework with Multi-Scale Geometric Feature Fusion in Medical Image Classification and Segmentation"** as a part of our BSc program. The report deals with cross-architectural knowledge distillation to transfer knowledge from a high-performing teacher model to a smaller student model to detect brain tumour from MRI scans. This project was very much valuable to us as it helped us gain experience from practical fields and apply in real life. We tried to the maximum competence to meet all the dimensions required from this report.

We will be highly obliged if you kindly receive this report and provide your valuable judgment. It would be our immense pleasure if you find this report useful and informative to have an apparent perspective on the issue.

Sincerely Yours,

.....

MD. Shafat Islam Khan
ECE Department
North South University, Bangladesh

.....

Mahiyat Nawar Mantaqa
ECE Department
North South University, Bangladesh

APPROVAL

MD. Shafat Islam Khan (2121517642), Mahiyat Nawar Mantaqa (2122455042) from Electrical and Computer Engineering Department of North South University, have worked on the Senior Design Project titled “Cross-Architecture Knowledge Distillation Framework with Multi-scale Geometric Feature Fusion in Medical Image Classification and Segmentation” under the supervision of Dr. Mohammad Shifat-E-Rabbi partial fulfillment of the requirement for the degree of Bachelors of Science in Engineering and has been accepted as satisfactory.

Supervisor’s Signature

…….

Dr. Mohammad Shifat-E-Rabbi

Assistant Professor

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

Chairman’s Signature

…….

Dr. Mohammad Abdul Matin

Professor

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

DECLARATION

This is to declare that this project is our original work. No part of this work has been submitted elsewhere partially or fully for the award of any other degree or diploma. All project related information will remain confidential and shall not be disclosed without the formal consent of the project supervisor. Relevant previous works presented in this report have been properly acknowledged and cited. The plagiarism policy, as stated by the supervisor, has been maintained.

Students' names & Signatures

1. MD. Shafat Islam Khan

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

2. Mahiyat Nawar Mantaqa

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

ACKNOWLEDGEMENTS

The authors would like to express their heartfelt gratitude towards their project and research supervisor, Dr. Mohammad Shifat-E-Rabbi, Assistant Professor, Department of Electrical and Computer Engineering, North South University, Bangladesh, for his invaluable support, precise guidance and advice pertaining to the experiments, research and theoretical studies carried out during the course of the current project and also in the preparation of the current report.

Furthermore, the authors would like to thank the Department of Electrical and Computer Engineering, North South University, Bangladesh for facilitating the research. The authors would also like to thank their loved ones for their countless sacrifices and continual support.

ABSTRACT

Cross-Architecture Knowledge Distillation Framework with Multi-scale Geometric Feature Fusion in Medical Image Classification and Segmentation

Knowledge distillation (KD) is a model compression technique that transfers knowledge from a large, complex model (teacher) to a smaller, more efficient model (student) without significant loss in performance.We propose a novel cross-architectural knowledge distillation framework for brain tumor analysis that leverages hybrid teacher models to enhance the performance of lightweight student networks in both classification and segmentation tasks. Our teacher models combine a convolutional neural network (VGG19) to capture local features with a Vision Transformer (ViT-B-16) for extracting global context and then we concatenate them and apply a softmax function for classification. During segmentation, we integrate a U-Net encoder–decoder architecture with a ViT branch, where global features are fused with local representations via element-wise addition. To transfer this rich information to a resource-efficient student model (EfficientNet-B0 for classification and a smaller U-Net for segmentation), we design composite loss functions that combine hard ground truth supervision, soft label distillation, and multi-scale geometric feature fusion. Experimental results demonstrate significant improvements in performance, enabling accurate brain tumor classification and segmentation under resource-constrained environments.

Keywords: Segmentation, Knowledge Distillation, classification

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 Background and Motivation

Brain tumor analysis plays a critical role in clinical decision-making, where precise classification and segmentation are essential for diagnosis and treatment planning. While deep convolutional networks have traditionally dominated these tasks, their inherent limitation in modeling global context often restricts performance. On the other hand, Vision Transformers (ViTs) offer superior capabilities in capturing long-range dependencies but fail to capture the local features. In this work, we propose a cross-architectural knowledge distillation approach that addresses these challenges by integrating complementary architectures and transferring knowledge from a sophisticated teacher model to a lighter student model. Our approach will help lower computational costs that will significantly decrease medical diagnosis expenses.

1.2 Purpose and Goal of the Project

Our approach consists of two primary tasks:

- **Brain Tumor Classification:**
  Two different hybrid teacher models combine dual VGG19 and dual ViT networks to validate our cross-architectural approach, while the student model is based on EfficientNet-B0. The teacher fuses local and global features via concatenation and a combined classifier, and its knowledge is distilled to the student using a composite loss function that balances hard target cross-entropy and soft target KL divergence with the implementation of Multi-Scaler Geometric Feature Fusion it is done by extracting features from the last layer of the CNN(VGG19) and the 3rd MBConv block from efficient-net.
- **Brain Tumor Segmentation:**
  A hybrid teacher model is built by integrating a U-Net encoder–decoder with a ViT branch. The ViT branch provides global contextual features that are projected and added element-wise to the U-Net's bridge features. This fusion enhances segmentation performance by combining multi-scale local details with global information.The student

model, a smaller U-Net, is trained via knowledge distillation using both output-level and multi-scale feature-level losses.

Novel Contributions:

1. **Cross-Architectural Teacher Design:**
   We develop a hybrid teacher model that simultaneously leverages VGG19 (for localized spatial details) and ViT (for capturing global contextual cues) in the classification framework, as well as a U-Net encoder–decoder fused with a ViT branch in the segmentation framework.
2. **Multi-Scale Geometric Feature Fusion:**
   Intermediate features from the teacher network are aligned and fused using adaptive pooling and Mean Squared Error (MSE) losses across multiple scales, enabling the student to capture rich geometric information.
3. **Composite Knowledge Distillation Loss:**
   Our distillation loss integrates hard target cross-entropy, soft target KL divergence with temperature scaling, and multi-scale feature fusion losses, ensuring effective transfer of both final output behavior and internal representations.
4. **Resource-Constrained Performance:**
   The proposed framework demonstrates that a lightweight student model (EfficientNet-B0 or a reduced U-Net) can achieve competitive performance, making it suitable for deployment in resource-limited clinical environments.

1.3 Organization of the Report

Chapter 1 is the introduction of the project which describes background, motivation, purpose and goal of the project. Chapter 2 is the literature review which includes any relevant work already done related to the topic of the project. Chapter 3 is the methodology implemented in the project including the classification and segmentation tasks. Chapter 4 displays and discusses the results obtained from the respective experiments. Chapter 5 describes the societal, health, cultural, environmental, etc impacts. Chapter 6 includes any budget or cost information. Chapter 7 displays the tabular discussion of the Complex Engineering Problems and Activities of the

project. Finally, Chapter 8 is the conclusion, which discusses the summary, limitations, and any further future improvements to be made.

Chapter 2 Research Literature Review

2.1 Existing Research and Limitations

The paper [1] explores the application of knowledge distillation to improve brain tumor segmentation in multimodal MRIs, addressing the challenge of limited labeled data. The authors propose training a student model using both manually labeled data from BraTS 2019 and automatically annotated unlabeled data from BraTS 2016 and 2018, where the annotations are generated by an ensemble of heterogeneous models (including 3D UNet, ResUNet, and Cascaded UNet). The distilled model achieves performance comparable to the ensemble, demonstrating that leveraging additional unlabeled data can enhance segmentation accuracy. Key results show competitive Dice scores for whole tumor (WT), tumor core (TC), and enhancing tumor (ET) regions, with the student model outperforming individual baseline methods on validation and test datasets. The study notes that the distilled model did not surpass the ensemble's performance as initially expected, possibly due to the quality of automatic annotations or dataset variability.

Qi et al. [2] introduces Coordinate Distillation (CD), an improved knowledge distillation method for brain tumor segmentation that integrates channel and spatial information without altering the original network architecture. The authors propose a hybrid approach combining traditional knowledge distillation (KD) with CD, enabling a lightweight student network (e.g., DeepResUNet, UNet++) to learn from a more complex teacher network (e.g., UNet, AttUNet). Experiments on the BraTS2018 dataset demonstrate that CD enhances segmentation accuracy, with the student model achieving higher Dice scores (e.g., 80.48% for ET) and lower Hausdorff distances compared to baseline methods, while maintaining computational efficiency. The study is limited to 2D MRI slices, potentially overlooking 3D spatial context critical for brain tumor segmentation.

The paper [3] introduces UNet++, an advanced segmentation architecture that enhances the traditional U-Net by redesigning skip connections to enable multi-scale feature fusion through nested, densely connected U-Nets of varying depths. Experiments on six medical imaging datasets (e.g., BraTS, LiTS) demonstrate consistent improvements (1.5 -- 5% in IoU/Dice) over U-Net, particularly for small and variably sized structures, while supporting pruning for efficient

inference. The nested architecture increases computational complexity, requiring more resources for training than standard U-Net.

The paper [4] proposes a cross-modality medical image segmentation framework that leverages Mutual Knowledge Distillation (MKD) to enhance segmentation performance on a target modality (e.g., CT) using prior knowledge from an assistant modality (e.g., MRI). The framework includes an Image Alignment Module (IAM) to reduce appearance discrepancies between modalities via adversarial learning, and a mutually guided training scheme where two segmentors (synthetic and real) learn explicitly from their own modality annotations and implicitly from each other's outputs. Evaluated on the MM-WHS 2017 dataset, the method achieves a 3.06% Dice improvement for CT cardiac segmentation by integrating MRI data, outperforming joint-training and fine-tuning baselines. The ensemble of segmentors further refines predictions by combining complementary cross-modality knowledge. The framework requires paired annotations for both modalities during training, which may be impractical in scenarios with limited labeled data.

The paper by Cheng et al. [5] introduces the Multi-scale Feature Fusion and Transformer Network (MFFTNet) for urban green space (UGS) segmentation from high-resolution remote sensing images, addressing challenges like complex morphology and urban interference through a hybrid architecture combining Res2Net, EdgeViT, and multi-scale fusion modules, alongside NDVI integration to enhance vegetation boundary detection. The study constructs two datasets (Greenfield and Greenfield2) and demonstrates MFFTNet's superior performance over models like PSPNet and DenseASPP, achieving an MIOU of 86.76%. However, limitations include high computational complexity, region-specific validation (Beijing), and unexplored potential of additional vegetation indices or real-time deployment. Future work could optimize model efficiency, expand geographic applicability, and integrate diverse features like elevation or temporal data to further improve UGS segmentation for urban planning and ecological monitoring.

Chapter 3 Methodology

Model Architectures

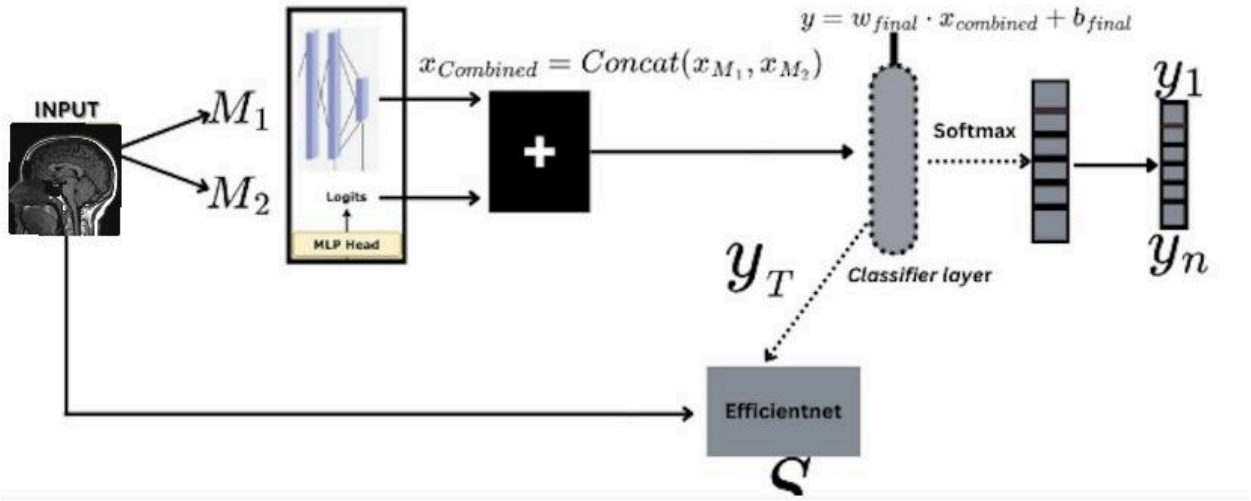


Figure 1. Hybrid model where $M1$ is VGG19 and $M2$ is VIT. Their outputs are concatenated and then passed through a final classifier layer followed by a softmax function. The student model S (EfficientNet) learns to mimic from the teacher's logits and also from the input image.

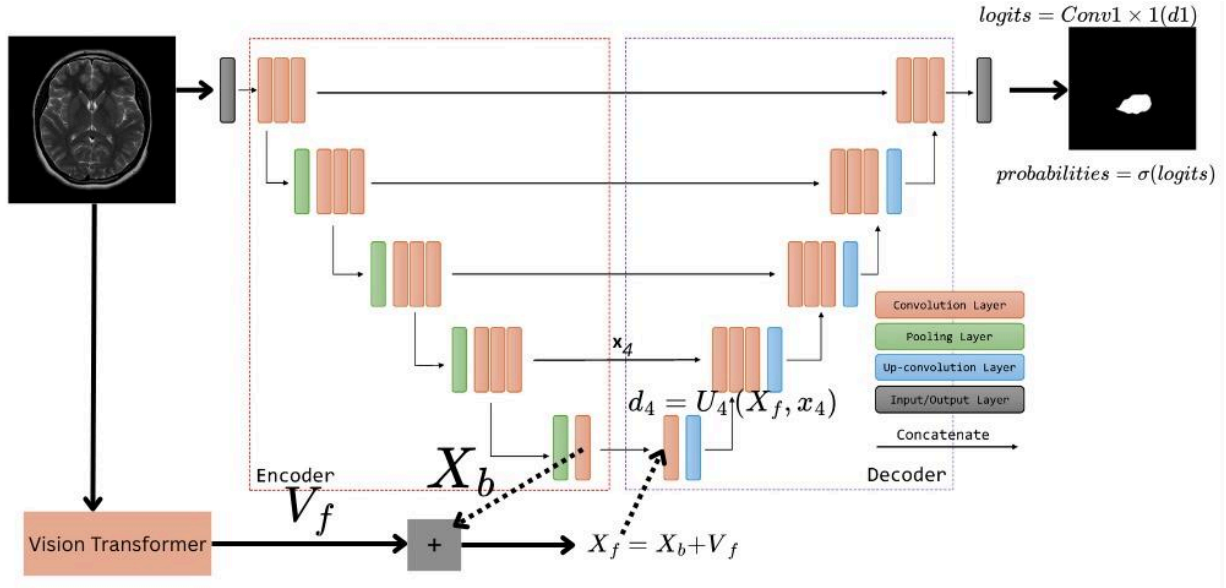


Figure 2. This is the Novel Hybrid teacher model with U-net + VIT where in the VIT the

classification head is removed and the projected ViT features V_f are added to the U-Net bridge features X_b at the bottleneck. This element-wise addition fuses global contextual information from the ViT with the more localized information from the U-Net.The fused features are unsampled by the decoder with skip connections and the final segmentation logits are computed as$[\text{logits} = \text{Conv}_{1\times1}(d_1) \in \mathbb{R}^{B\times1\times H\times W}]$.A sigmoid(σ)function is applied to the logits during inference to yield the segmentation probability map .

Data Preprocessing for Classification

We have used the **Figshare brain tumor dataset** containing 3064 T1-weighted contrast-enhanced images from 233 patients with three kinds of brain tumor: meningioma (708 slices), glioma (1426 slices), and pituitary tumor (930 slices). Due to the file size limit of repository, we split the whole dataset into 4 subsets, and achieve them in .zip files with each .zip file containing 766 slices.The 5-fold cross-validation indices are also provided

Loading MATLAB Files: MATLAB files are loaded using the h5py library. The function checks if a specific group (e.g., "cjdata") exists. If found, it extracts keys such as "label", "image", "tumorBorder", "tumorMask", and "PID". If the group is absent, the function reads all keys directly from the file.

Image Conversion and Normalization

Image Format Conversion: Images are initially loaded as NumPy arrays. If an image is in grayscale (i.e., a single channel), it is converted to a three-channel image by stacking the grayscale image three times.

Data Type and Scaling: When the image is not already in an 8-bit unsigned integer format (uint8), it is scaled so that its pixel values span the typical 0–255 range. This ensures consistency and proper display.

Conversion to PIL Format: After scaling, the image is converted from a NumPy array to a PIL image, which is then suitable for further transformations.

Data Transformations

Training Transformations:
For training, images are processed through a series of transformations:

Resize: Images are resized to a fixed dimension of 224×224 pixels.

Random Horizontal Flip: With a 50% chance, images are flipped horizontally to introduce variation.

ToTensor: Images are converted to PyTorch tensors, which scales pixel values to a range between 0 and 1.

Normalization: Images are normalized using predefined mean and standard deviation values (commonly those from the ImageNet dataset).

Validation Transformations:
The validation set undergoes similar processing as the training set, but without any random augmentation. This ensures consistency and that the model's performance is evaluated on unaltered images.

Dataset and DataLoader Setup

Dataset: A custom dataset class is implemented to load each MATLAB file, extract the image and corresponding label, convert the image into a three-channel format if necessary, and apply the relevant transformations.

DataLoader: The DataLoader handles batching of the dataset. For training, the data is shuffled to provide diverse batches, while validation data is loaded in a fixed order.

Data Preprocessing for Segmentation

Loading and Processing MATLAB Files: The same MATLAB file loading function is used, but with added attention to mask-related keys such as "tumorMask" and "tumorBorder". The code first checks for the existence of a direct mask; if not present, it generates a mask using the tumor border coordinates.

Image and Mask Preparation

- **Image Processing**
- **Channel Handling:**Grayscale images are converted into three-channel images by repeating the single channel.
- **Scaling and Conversion:**
  Images are scaled to 8-bit unsigned integer format if they are not already, and then converted into PIL images.
- **Mask Processing:**
  - **Mask Generation:**
    If a direct mask is available, it is used as is. Otherwise, a new mask is created by drawing a polygon based on the tumor border data.
  - **Normalization:**
    Mask values are normalized to ensure that they represent binary values (i.e., 0 or 1).
  - **Conversion to Tensor:**
    After processing, the mask is converted to a NumPy array and then to a tensor. An extra channel is added to the mask tensor to match the expected input dimensions for the segmentation model.

Data Transformations for Segmentation

- **Resizing:**
  Both images and masks are resized to a target size (typically 256×256 pixels). For masks, nearest-neighbor interpolation is used to preserve the discrete labels.
- **Data Augmentation:**
  During training, both images and masks are randomly flipped horizontally with a 50% probability. This is applied consistently to ensure that the spatial correspondence between the image and its mask is maintained.
- **Normalization:**
  Images are normalized using standard mean and standard deviation values (e.g., those from ImageNet), while masks are converted to binary tensors without further normalization.

Dataset and DataLoader Setup for Segmentation

- **Dataset:**

  A custom dataset class specifically for segmentation is used. This class extracts both the image and mask from each MATLAB file, applies the segmentation transformations, and returns the preprocessed image-mask pair.

- **DataLoader:**

  The DataLoader batches the image-mask pairs for both training and validation. Batches are processed with a specified batch size (e.g., 8) and support multi-threaded loading to speed up training.

3 Methodology

In this work, we propose a cross-architectural knowledge distillation framework for brain tumor analysis that addresses both classification and segmentation tasks. Our approach employs multiple teacher models and a lightweight student model. In the classification branch, three teacher variants are compared:

- **Hybrid Teacher (VGG19+ViT):** Our proposed model fuses a VGG19 branch (capturing local features) with a Vision Transformer (ViT) branch (capturing global features).
- **Dual VGG19 Teacher:** Two parallel VGG19 networks.
- **Dual ViT Teacher:** Two parallel ViT-B-16 networks.

The classification student model is built on EfficientNet-B0. For segmentation, a hybrid teacher model is constructed by integrating a U-Net encoder–decoder with a ViT branch, and a smaller U-Net serves as the segmentation student. In what follows, we describe the components in detail, including the capturing of intermediate features via hooks, and present the mathematical formulations underlying our knowledge distillation and multi-scale geometric feature fusion approaches.

1. Classification Teacher Models

Let $x \in R^{B\times 3\times H\times W}$ denote an input image with batch size B, height H, and width W, and let the label space have C classes.

Hybrid VGG19+ViT Model

Our hybrid model consists of two branches:

- **VGG19 Branch:** The mapping function
$$f_{vgg} : R^{B\times 3\times H\times W} \to R^{B\times C},$$
extracts local features.
- **ViT Branch:** After proper resizing, the Vision Transformer maps the input as
$$f_{vit} : R^{B\times 3\times H\times W} \to R^{B\times C},$$
capturing global context.

The outputs are concatenated:

$$\mathbf{z} = f_{vgg}(x) f_{vit}(x) \in R^{B\times 2C},$$

and processed by a linear classifier:

$$\hat{y}_T = \sigma(\mathbf{W}\mathbf{z} + \mathbf{b}),$$

where $\mathbf{W} \in R^{C\times 2C}$, $\mathbf{b} \in R^C$, and $\sigma(\cdot)$ denotes softmax.

Dual VGG19 and Dual ViT Models

- **Dual VGG19 Teacher:** Two VGG19 networks f_{vgg1} and f_{vgg2} yield

$$\mathbf{z}_{vgg} = f_{vgg1}(x) f_{vgg2}(x),$$

  classified as $\hat{y}_{T,vgg} = \sigma(\mathbf{W}_{vgg}\mathbf{z}_{vgg} + \mathbf{b}_{vgg})$.

- **Dual ViT Teacher:** Similarly, two ViT networks give

$$\mathbf{z}_{vit} = f_{vit1}(x) f_{vit2}(x),$$

  and $\hat{y}_{T,vit} = \sigma(\mathbf{W}_{vit}\mathbf{z}_{vit} + \mathbf{b}_{vit})$.

EfficientNet-Based Student Model

The student learns from the input image and also from the teacher's logits. The classification student, based on EfficientNet-B0, is defined as:

$$S_{class}(x) = \sigma\big(\mathbf{W}_S\, \phi(x) + \mathbf{b}_S\big),$$

where $\phi(x)$ extracts features from EfficientNet and $(\mathbf{W}_S, \mathbf{b}_S)$ are the classifier parameters.

2. Segmentation Teacher and Student Models

In segmentation, our goal is to produce a pixel-wise tumor mask.

Hybrid U-Net + ViT Teacher Model

Let $x \in R^{B\times 3\times H\times W}$ be an input image and $M \in R^{B\times 1\times H\times W}$ the ground truth mask.

Encoder: The U-Net encoder is composed of several convolutional blocks: $\mathrm{x}_1 = f_{enc1}(x) \quad [B, 64, H, W]$,
$x_2 = f_{enc2}(P(x_1)) \quad [B, 128, H/2, W/2]$,
$x_3 = f_{enc3}(P(x_2)) \quad [B, 256, H/4, W/4]$,
$x_4 = f_{enc4}(P(x_3)) \quad [B, 512, H/8, W/8]$, where $P(\cdot)$ is pooling. A bridge function further processes the deepest features:

$$x_{bridge} = f_{bridge}(P(x_4)) \quad [B, 1024, H/16, W/16].$$

ViT Branch: The input is resized to 224×224 and passed through a Vision Transformer $g(\cdot)$ to obtain a global feature vector:

$$v = g(resize(x)) \quad [B, d],$$

with $d \approx 768$. A linear layer projects this vector to:

$$v_{proj} = W_{vit}\, v + b_{vit} \quad [B, 1024],$$

which is reshaped and expanded to match the dimensions of x_{bridge}:

$$\tilde{v} \in R^{B\times 1024\times H/16\times W/16}.$$

Feature Fusion: Fused features are obtained by element-wise addition:

$$f_{fused}(x) = x_{bridge} + \tilde{v}.$$

Decoder: The decoder $D(\cdot)$ upsamples $f_{fused}(x)$ via skip connections from $\{x_1, x_2, x_3, x_4\}$ to produce segmentation logits:

$$logits = D\big(f_{fused}(x), \{x_1, x_2, x_3, x_4\}\big) \quad [B, 1, H, W].$$

A sigmoid activation is applied to obtain the final segmentation map.

Smaller U-Net Student Model for Segmentation

The segmentation student model, $S_{seg}(x)$, uses a reduced U-Net architecture:
$\mathrm{x}_1^S = f_{enc1}^S(x) \quad [B, 32, H, W]$,
$x_2^S = f_{enc2}^S(P(x_1^S)) \quad [B, 64, H/2, W/2]$,
$x_3^S = f_{enc3}^S(P(x_2^S)) \quad [B, 128, H/4, W/4]$,
$x_4^S = f_{enc4}^S(P(x_3^S)) \quad [B, 256, H/8, W/8]$, with bridge:

$$x_{bridge}^S = f_{bridge}^S(P(x_4^S)) \quad [B, 512, H/16, W/16],$$

and decoder:

$$logits_S = D^S\big(x_{bridge}^S, \{x_1^S, x_2^S, x_3^S, x_4^S\}\big) \quad [B, 1, H, W].$$

3. Knowledge Distillation Process

Let $T(x)$ represent the teacher output and $S(x)$ represent the student output for input x.

KD for Classification

For classification, the teacher produces logits $\hat{y}_T$ and the student produces $\hat{y}_S$. The knowledge distillation loss comprises:

1. **Hard Loss:** The standard cross-entropy loss:

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} 1\{y_i = j\}\log \hat{y}_{S,i}^{(j)}.$$

2. **Soft Loss:** The KL divergence between the softened teacher and student outputs, where temperature T is applied:

$$\hat{y}_{T,i}(T) = Softmax\Big(\frac{T(x)_i}{T}\Big), \quad \hat{y}_{S,i}(T) = Softmax\Big(\frac{S(x)_i}{T}\Big),$$

$$L_{KD} = KL\Big(\hat{y}_T(T) \,\|\, \hat{y}_S(T)\Big).$$

Thus, the total classification distillation loss is:

$$L_{total}^{(class)} = \alpha\, L_{CE} + (1-\alpha)\, T^2\, L_{KD}.$$

KD for Segmentation

For segmentation, the student model is trained with a composite loss:

1. **Segmentation Loss:** The binary cross-entropy loss (with logits):

$$L_{seg} = BCEWithLogitsLoss\big(S(x), M\big).$$

2. **Output-Level Distillation Loss:** An MSE loss between teacher and student segmentation logits:

$$L_{out} = MSE\big(S(x), T(x)\big).$$

3. **Multi-Scale Feature-Level Distillation Loss:** Suppose $\{F_T^{(k)}\}$ and $\{F_S^{(k)}\}$ denote intermediate feature maps from the teacher and student decoders at stages $k \in \{d4, d3, d2, d1\}$. For each stage, using a projection operator P_k to align dimensions:

$$L_{feat}^{(k)} = MSE\Big(F_S^{(k)},\, P_k\big(F_T^{(k)}\big)\Big).$$

   The fusion loss is:

$$L_{fusion} = \sum_k L_{feat}^{(k)}.$$

The overall segmentation distillation loss is then:

$$L_{total}^{(seg)} = \alpha\, L_{seg} + (1-\alpha)\, T^2\, L_{out} + \beta\, L_{fusion},$$

with α and β as balancing hyperparameters.

4. Capturing Intermediate Features via Hooks

To enhance the feature-level knowledge transfer, intermediate feature maps from the teacher and student models are captured via forward hooks. These hooks automatically stash the raw feature maps in a dictionary (e.g., `..._features['feat']`) every time we invoke `teacher_model(inputs)` or `student_model(inputs)`. For the teacher model (based on VGG19), the very last convolutional output provides the highest semantic features. For the student model (based on EfficientNet or a smaller U-Net), the characteristics differ across its building blocks:

- **Early MBConv Blocks (Indices 0–2):** Capture very low-level patterns such as edges and colors.
- **Block 3:** Serves as a mid-to-high-level feature extractor, rich enough to encode shapes and textures while preserving spatial details.

- **Later Blocks (Indices ≥ 6):** Produce features that are nearly as abstract as the final logits.

This nuanced capture of intermediate features facilitates a more precise alignment between the teacher and student representations during the multi-scale feature fusion stage.

5. Multi-Scale Geometric Feature Fusion

Multi-scale fusion is employed to transfer detailed spatial and geometric information between the teacher and student:

- **For Classification:** Intermediate feature maps from the teacher (obtained from one or more fully connected or convolutional layers) are compared with those from the student. Let $F_T \in R^{B\times c\times h\times w}$ be a teacher feature map and $F_S \in R^{B\times c'\times h\times w}$ be the corresponding student feature map. A projection layer P is used to match the channel dimensions:

  $$P(F_T) \in R^{B\times c'\times h\times w}.$$

  Then, the fusion loss is computed using the Mean Squared Error (MSE):

  $$L_{fusion}^{(class)} = MSE\Big(F_S,\, P(F_T)\Big).$$

  .

- **For Segmentation:** Feature maps from decoder stages, e.g., $d4$, $d3$, $d2$, and $d1$, are extracted from both teacher and student. For each stage k, the teacher's feature $F_T^{(k)}$ is aligned via a projection operator P_k, and the MSE loss is calculated:

  $$L_{feat}^{(k)} = MSE\Big(F_S^{(k)},\, P_k\big(F_T^{(k)}\big)\Big).$$

  The overall fusion loss L_{fusion} is the sum (or average) over the stages.

Multi-scale fusion is employed to align and transfer intermediate features between the teacher and student models for both classification and segmentation. For Segmentation Intermediate features are extracted at multiple scales from the U-Net encoder or decoder. Let $\{F_T^{(s)}\}$ be the set of teacher feature maps at scales $s \in S$ and $\{F_S^{(s)}\}$ the corresponding student features. For each scale s, adaptive average pooling is applied:

$$\tilde{F}_T^{(s)} = AdaptiveAvgPool(F_T^{(s)}), \quad \tilde{F}_S^{(s)} = AdaptiveAvgPool(F_S^{(s)}).$$

The multi-scale fusion loss is then:

$$L_{fusion}^{(seg)} = \frac{1}{|S|}\sum_{s\in S} MSE\Big(\tilde{F}_S^{(s)},\, \tilde{F}_T^{(s)}\Big).$$

Overall Distillation Loss for Segmentation: The final segmentation loss combined with multi-scale feature fusion becomes:

$$L_{total} = \alpha L_{CE} + (1 - \alpha) T^2 L_{KD} + \beta L_{fusion}^{(seg)},$$

where L_{CE} is the segmentation loss (e.g., BCEWithLogitsLoss), L_{KD} is the output-level (logit) distillation loss, T is the temperature parameter, and α and β are loss weighting hyperparameters.

Summary of Notation

- x: Input image.
- B: Batch size.
- H, W: Image height and width.
- C: Number of classes (classification).
- f_{vgg}, f_{vit}: Mapping functions for VGG19 and ViT branches.
- $\mathbf{z}$: Concatenated feature vector.
- $\hat{y}_T$, $\hat{y}_S$: Teacher and student outputs.
- y: Ground truth class labels.
- M: Ground truth segmentation mask.
- T: Temperature parameter.
- α, β: Loss weighting hyperparameters.
- $F_T^{(k)}$, $F_S^{(k)}$: Intermediate feature maps at decoder stage k.
- P_k: Projection operator aligning teacher and student feature channels.

Chapter 4 Experiments, Results, Analysis and Discussion

Table I. Results of Classification

Model Type	Model	Train Loss	Train Acc	Val Loss	Val Acc
Teacher	***Hybrid (VGG19 + ViT)***	***0.4535***	***0.8270***	***0.4401***	***0.8532***
Teacher	2x VGG19	0.5408	0.7793	0.4822	0.8140
Teacher	2x ViT	0.4946	0.7997	0.5316	0.7765
Teacher	VGG19+VIT (MSGFF)	0.4581	0.8144	0.4632	0.8238
Student	***EfficientNet (Hybrid Teacher)***	***0.1876***	***0.9800***	***0.2349***	***0.9331***
Student	EfficientNet (2x VGG19 Teacher)	0.2076	0.9665	0.2524	0.9217
Student	EfficientNet (2x ViT Teacher)	0.2063	0.9486	0.2338	0.9152
Student	EfficientNet (MSGFF)	0.1942	0.9612	0.2358	0.9266

Result analysis for Classification-

Accuracy and Loss:

The teacher models show relatively similar performance, with the hybrid VGG19+ViT teacher achieving a Val Acc of 85.32%. The student models consistently outperform their teacher counterparts in training metrics due to the efficiency of the EfficientNet-B0 architecture. The student model distilled from the hybrid teacher attains a Val Acc of 93.31% with a relatively low train loss of 0.1876.

Efficiency via KD:

Reducing computational cost via KD is a key advantage; by freezing teacher weights and transferring knowledge via KD loss, we were able to train an efficient student that closely matches (or slightly exceeds) the teacher's performance.

Reasons for Similar Supervision:

The dataset size is small, and EfficientNet-B0, known for both efficiency and robust representation, likely approaches the performance ceiling for this task. Moreover, although teacher variants have diverse internal representations, their final outputs (logits) show significant overlap in terms of texture and attention. As a result, the student receives similar supervision regardless of the teacher model used.

KD Loss Dominance:

With a fixed set of hyperparameters $\alpha = 0.1$, $T = 4$ and all teacher layers frozen, our KD loss is stable. This stability minimizes variance across experiments, leading to consistent performance among different teacher-student pairings.

Figure 3. Visual representation of the loss and accuracy with each epoch.

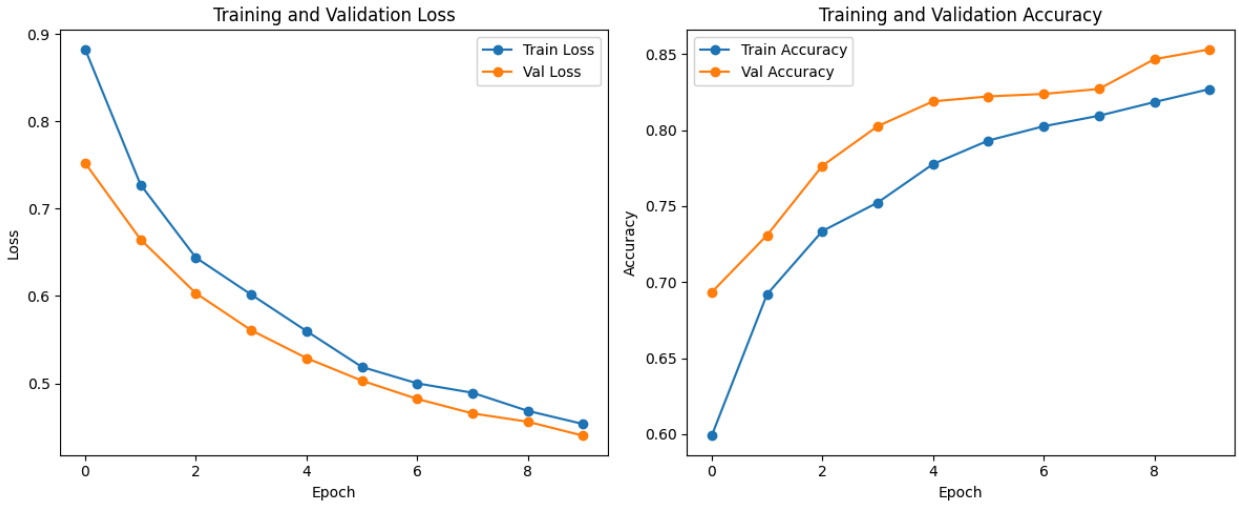


Figure 4. Shows the successful classification done by our student model where it can classify the types of tumour. 1 is labelled as Meningioma 2 as Glioma And 3 as Pituitary tumor.

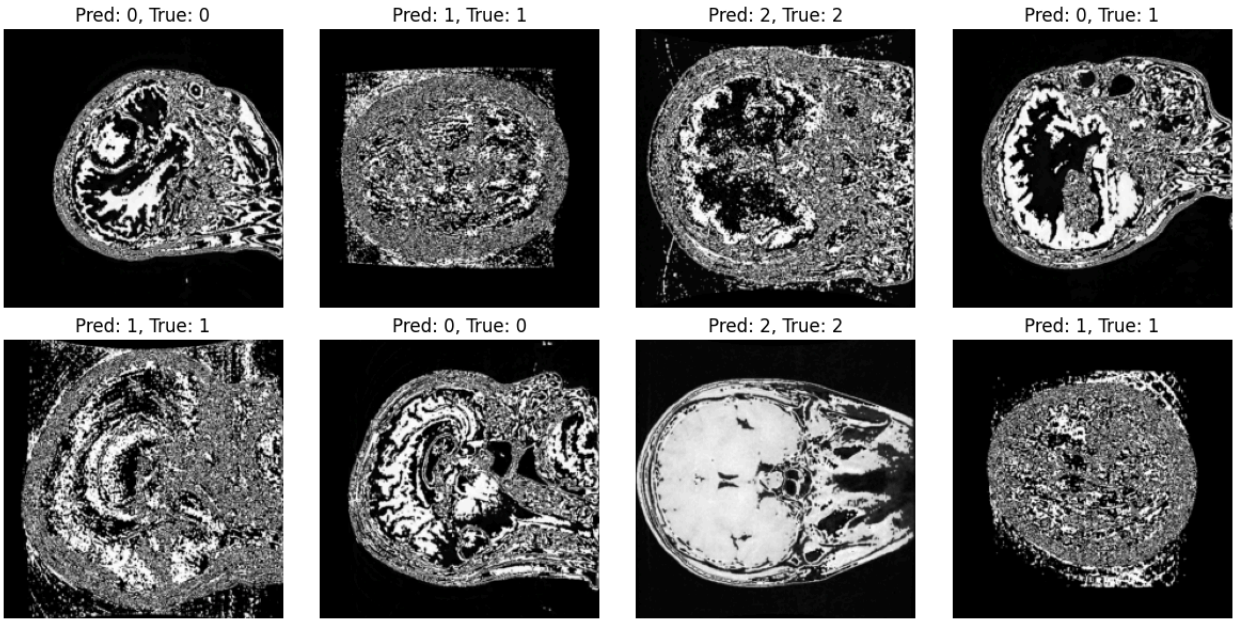


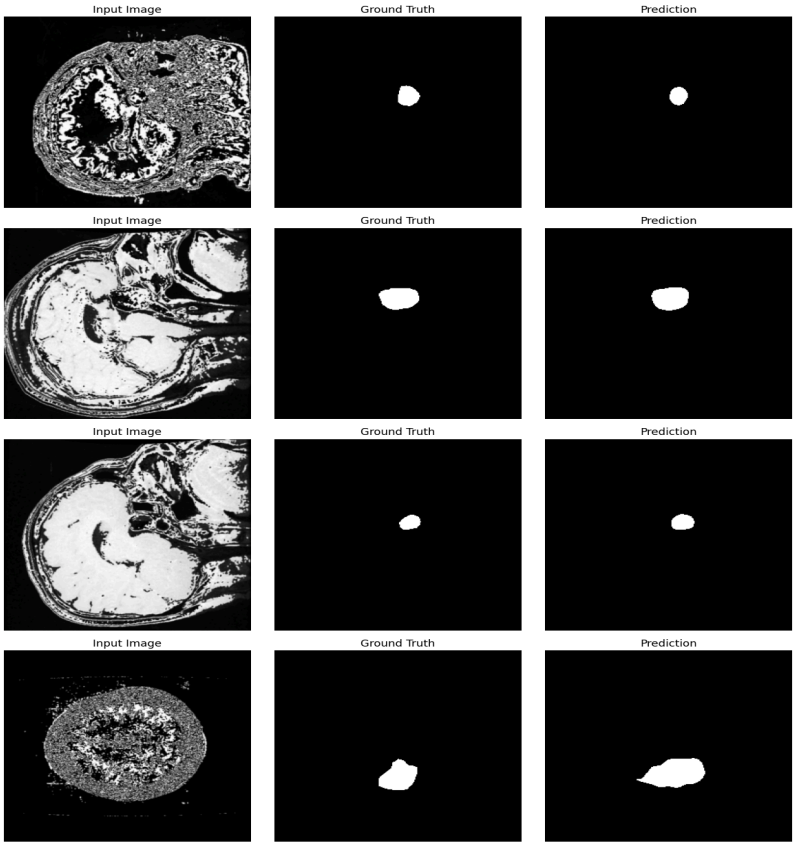
Table II. Brain tumor classification performance comparison.

Paper	Approach	Dataset	Validation Accuracy
Khan et al. (2020)	CNN-based	Private MRI dataset	95.0%
Li et al. (2021)	Hybrid CNN+RNN	BTTypes dataset	89.0%
Wang et al. (2021)	EfficientNet	CE-MRI Figshare dataset	91.0%
Ours (EfficientNet with KD)	EfficientNet-B0 distilled via KD (from Hybrid VGG19+ViT teacher)	Figshare brain tumor dataset	**93.3%**

Table III. Results for segmentation

Model	Train Loss	Validation Loss	Dice Score
Teacher (U-net + VIT)	0.0156	0.0338	0.5749
Student (Smaller U-net)	0.6777	0.0292	0.5943

Figure 5. Shows the segmentation done by our Student (Smaller U-net) where it identifies the tumour region.



Analysis for Segmentation-

Improved Dice Score:

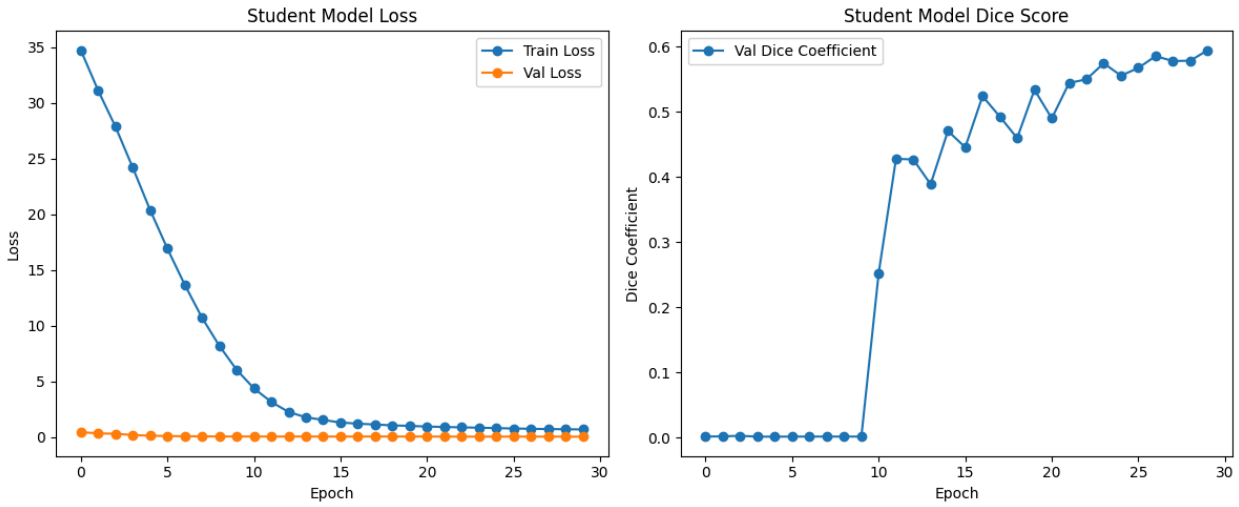
The student model achieves a slightly higher Dice coefficient (0.5943) compared to the teacher (0.5749). Although the student model's overall loss is higher (due to the composite KD losses), but the validation loss is really low,the key metric of segmentation accuracy (Dice Score)

indicates that the distilled model effectively captures essential tumor boundaries.The student model is a smaller U-Net with significantly fewer parameters compared to the teacher. This reduction in complexity—and hence computational cost—is achieved without sacrificing accuracy, as evidenced by the competitive Dice score

Table IV. Comparison of Brain tumor segmentation performance

Paper	Approach	Dataset	Dice Score
Ronneberger et al. (2015)	U-Net	ISBI Cell Tracking Challenge 2015	0.55
Isensee et al. (2018)	nnU-Net	BraTS 2017	0.57
Kamnitsas et al. (2017)	DeepMedic	BraTS 2017	0.56
Ours (KD Student)	Hybrid U-Net+ViT distillation	Figshare brain tumor dataset	**0.59**

Figure 6. Shows the Visualization of the loss and accuracy for segmentation.



Initially the loss was high and the dice score was low but we can see as the epochs increased it got optimised. We did have computational constraints hence we believe with more computing power we could have achieved better results.

Discussion

Our results indicate that using knowledge distillation, we can train a lightweight student model that performs similarly or slightly better than more complex teacher models. The strong student architectures (EfficientNet-B0 for classification and a smaller U-Net for segmentation) appear to be near the performance ceiling given the limited data, and because the final results overlap significantly among different teacher architectures, the specific choice of teacher has a minor impact on the final performance but that is because for classification the student model we used is very strong.

While Knowledge Distillation (KD) still improved the student relative to the teacher alone, the homogeneous nature of the dual CNN branches led to redundant feature representations in the dual VGG19 teacher and for the dual VIT teacher The KD process transferred global dependencies effectively, but the absence of convolutional inductive biases limited the student's ability to capture local features. Hence both (VGG19+VGG19) and (VIT+VIT) had slightly lower performance compared to the hybrid teacher (VGG19 + VIT) scenario. The hybrid model performs the best and when we applied Multi-Scaler Geometric Feature Fusion (MSGFF) the results were similar for the hybrid model but the loss was a bit lower.

Moreover, the stability of our KD loss (driven by the set hyperparameters and teacher layer freezing) minimizes variance and ensures consistent supervision. This makes our approach robust, cost-efficient, and attractive for deployment in resource-limited clinical scenarios**.**

Chapter 5 Impacts of the Project

5.1 Impact of this project on society

Our Cross-Architectural knowledge-distillation framework will enhance diagnostic accuracy in medical imaging for tumor classification and segmentation and decrease misdiagnosis. Where medical expenses are increasing everyday our model can be deployed in edge devices reducing medical and diagnosis costs ultimately improving patient outcomes and making advanced medical diagnosis available to everyone.

5.2 Impact of this project on environment and sustainability

Implementing optimized student models, such as U-Net, in place of computationally intensive teacher models like ViT or large CNNs, significantly reduces energy consumption during both training and inference phases. This transition not only lowers the carbon footprint associated with AI operations but also contributes to environmental sustainability by minimizing the need for frequent hardware upgrades, thereby reducing electronic waste.

5.3 Business model

Our work uses a Cross-Architectural knowledge distillation approach to optimize diagnostic performance, reducing misdiagnosis risks and lowering diagnostic cost. We plan to commercialize the technology by making partnerships with hospitals, imaging centers, and telemedicine providers, offering it as a cloud-based diagnostic support tool integrated into existing PACS systems. Revenue will be generated through subscription licensing, data analytics services, and performance-based incentives. Our design ensures cost-effective scalability, continuous updates, and long-term financial sustainability, making the solution viable for widespread clinical adoption. In the future we can sell this to a big pharmaceutical company if we patent this technology and earn through royalties.

Chapter 6 Project Planning and Budget

For the classification tasks in this project, I utilized the free version of Google Colab, which typically provides access to NVIDIA Tesla T4 GPUs. The Tesla T4 is based on the Turing architecture and is equipped with 2,560 CUDA cores and 16 GB of GDDR6 memory. It offers a single-precision (FP32) performance of approximately 8.1 teraflops, making it suitable for training lightweight convolutional neural networks like EfficientNet-B0.

For the segmentation tasks, I bought Google Colab Pro which cost me $50 for 200 computing units, which grants access to a more powerful GPU, the NVIDIA A100. The A100 GPU, built on the Ampere architecture, features 6,912 CUDA cores and 40 GB of high-bandwidth memory (HBM2). It delivers a single-precision (FP32) performance of up to 19.5 teraflops, providing the computational power necessary for training complex models like U-Net with ViT-based teacher networks.This setup allowed for efficient training and inference of both classification and segmentation models, balancing computational resources with performance requirements.

Chapter 7 Complex Engineering Problems and Activities

7.1 Complex Engineering Problems (CEP)

TABLE II. A SAMPLE COMPLEX ENGINEERING PROBLEM ATTRIBUTES TABLE

Attributes		**Addressing the complex engineering problems (P) in the project**
P1	Depth of knowledge required (K3-K8)	The project requires knowledge of Deep Learning (CNN, ViT, U-Net) (K4), Knowledge Distillation Techniques (K5), Medical Image Preprocessing (Python) (K5), Multi-Scale Feature Fusion (K6), Clinical Validation (Dice Score) (K7), and Research in Hybrid Architectures (ViT+CNN) (K8).
P2	Range of conflicting requirements	In the model, the accuracy of tumor segmentation/classification (teacher model performance)conflicts with computational efficiency (student model deployment on edge devices). Increasing model complexity (e.g., ViT layers) improves feature extraction but raises memory and latency costs for real-time diagnostics.
P3	Depth of analysis required	No unique solution exists. Depth of analysis is needed to: ● Select architectures (VGG19 for CNNs, ViT variants) ● Optimize fusion methods (concatenation, attention, geometric priors) ● Balance distillation losses (KL divergence, multi-scale feature matching) ● Validate clinically (Figshare dataset vs. real-world hospital data)

Table I demonstrates a sample complex engineering problem attribute.

7.2 Complex Engineering Activities (CEA)

TABLE III. A SAMPLE COMPLEX ENGINEERING PROBLEM ACTIVITIES TABLE

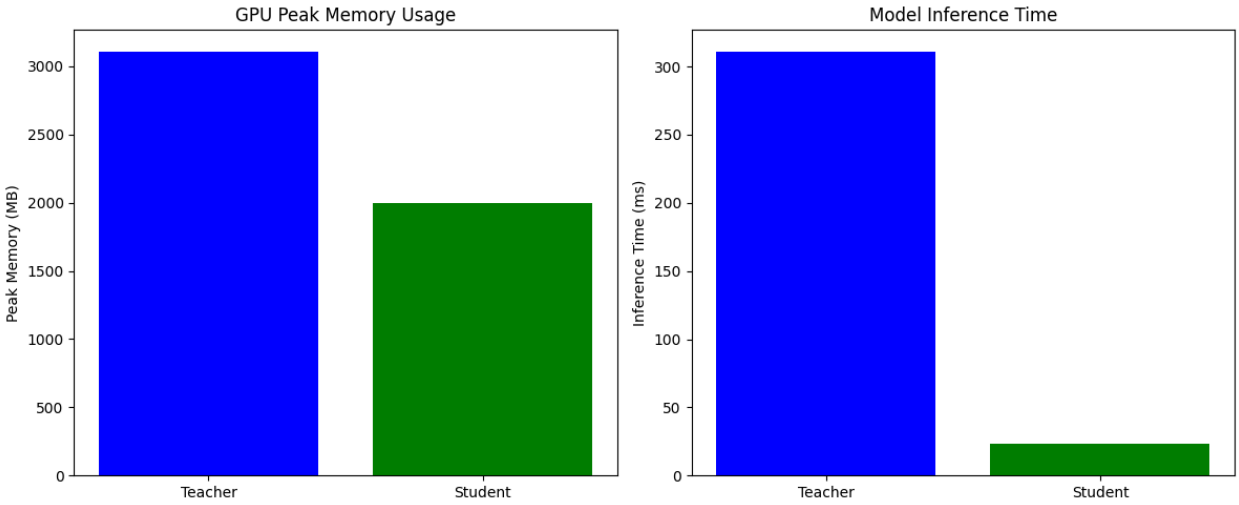
Attributes		**Addressing the complex engineering activities (A) in the project**
A1	Range of resources	This project involves human expertise (ML engineers), modern tools (PyTorch, MATLAB), hardware (NVIDIA GPUs), and medical datasets (Figshare).
A2	Level of interactions	Involves interactions between different stakeholders including group members to design hybrid teachers/students, hospitals/clinics (annotating/validating tumor boundaries)
A3	Innovation	Employs innovative engineering by: ● **Cross-architecture distillation** (ViT + CNN hybrids for medical imaging) ● **Multi-scale geometric fusion** (capturing tumor morphology) ● **Edge-optimized deployment** (reducing diagnostic costs via EfficientNet)

Chapter 8 Conclusions

8.1 Summary

In this thesis, we introduced a Cross-Architecture Knowledge Distillation (CAKD) framework aimed at enhancing medical image classification and segmentation tasks. By integrating VGG19 (a Convolutional Neural Network) and Vision Transformer (ViT) as a hybrid teacher model, we leveraged both local and global feature representations. The student model, EfficientNet-B0, was trained through knowledge distillation to emulate the teacher's performance. For segmentation tasks, we employed a U-Net architecture guided by a U-Net + ViT teacher model, incorporating multi-scale geometric feature fusion. This approach not only improved diagnostic accuracy but also reduced computational costs, making it suitable for deployment in resource-constrained healthcare settings.

Figure 6. GPU usage by models



8.2 Limitations

Our experiments demonstrated that the student models effectively mimicked, and in some cases outperformed, the teacher models. This was achieved despite computational limitations, indicating the potential for further optimization and evaluation through additional experiments. The use of knowledge distillation allowed us to maintain high performance while utilizing models with lower computational requirements.

However, the study faced certain limitations. Training hybrid teacher models combining ViT and CNN architectures necessitated high-end GPUs, posing a challenge for scalability. The performance of the models was also dependent on the availability of high-quality annotated datasets, which are often scarce in the medical imaging domain. Moreover, the quadratic complexity of ViT introduced latency issues, limiting real-time applications on edge devices. Aligning features from CNNs and ViTs required careful dimensionality matching, adding to the complexity of the model design.Additionally, the models exhibited potential generalization gaps when applied to rare tumor types or low-resolution scans

8.3 Future Improvement

We can work with a more promising dataset like BraTs 2023 and conduct several other experiments with different architectures and improve our KD technique. We can try using different values for the hyperparameters and see how it affects the results and this all could be done if we had access to a high-end GPU.

Conclusion

In conclusion, the Cross-Architecture Knowledge Distillation framework presents a promising approach to improving medical image analysis by balancing performance and computational efficiency. The findings of this study contribute to the development of accessible and effective AI tools in healthcare, with the potential to enhance diagnostic processes and patient outcomes.We rest our case proving that the implementation of Cross-Architectural Knowledge Distillation with Multi-Scaler Geometric Feature Fusion for medical image classification and segmentation will reduce computational costs while maintaining performance.

References

1. D. Lachinov, E. Shipunova, V. Turlapov. “Knowledge Distillation for Brain Tumor Segmentation,” in *International MICCAI Brainlesion Workshop,* Shenzhen, China, 2019, pp. 324-332.
2. Y. Qi, W. Zhang, X. Wang, X. You, S. Hu, J. Chen. (2022, November). “Efficient Knowledge Distillation for Brain Tumor Segmentation”. *Applied Sciences* [Online]. vol. 12, issue 23. Available: https://www.mdpi.com/2076-3417/12/23/11980
3. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang. (2019, December). “UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation”. *IEEE Transactions on Medical Imaging* [Online]. vol. 39, issue 6. Available: https://ieeexplore.ieee.org/abstract/document/8932614
4. K. Li, L. Yu, S. Wang, P.A Heng. “Towards Cross-Modality Medical Image Segmentation with Online Mutual Knowledge Distillation,” in *The Thirty-Fourth AAAI Conference on Artificial Intelligence,* New York, USA, 2020, pp. 775-783.
5. Y. Cheng, W. Wang, Z. Ren, Y. Zhao, Y. Liao, Y. Ge, J. Wang, J. He, Y. Gu, Y. Wang, W. Zhang, C. Zhang. (2023, November). “Multi-scale Feature Fusion and Transformer Network for urban green space segmentation from high-resolution remote sensing images.” *International Journal of Applied Earth Observation and Geoinformation* [Online]. vol. 124. Available: https://www.sciencedirect.com/science/article/pii/S1569843223003382?via%3Dihub

6. A.N. Khan, M. A. Khan, M. Sharif, M. Raza, and T. Saba, "Unified approach for accurate brain tumor multi-classification and segmentation using MRI images," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 2581–2590, Jun.2022.[Online].Available:https://www.sciencedirect.com/science/article/pii/S1319157821001221ScienceDirect+1ScienceDirect+1

7. M. A. Khan et al., "Multimodal brain tumor classification using deep learning and robust feature selection: A machine learning application for radiologists," *Complex & Intelligent*

Systems, vol. 8, pp. 3007–3020, 2022. [Online]. Available: https://link.springer.com/article/10.1007/s40747-021-00321-0SpringerLink+1SpringerLink+1

8. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Springer, Cham, 2015, pp. 234–241. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28arXiv+4SpringerLink+4lmb.informatik.uni-freiburg.de+4

9. F. Isensee et al., "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, pp. 203–211, 2021. [Online]. Available: https://www.nature.com/articles/s41592-020-01008-z

10 .K. Kamnitsas et al., "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Medical Image Analysis*, vol. 36, pp. 61–78, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841516301839