PUNITH_N_FINAl_REPORT_THEISIS_Ensemble_learning_v8.6.docx

by Punith N

A Review and Implementation of Ensemble Methods for Medical Image Segmentation in Computer Vision

Punith N

Final Thesis Report

FEB 2026

ABSTRACT

This thesis explores ensemble methods that combine models to improve medical image segmentation. It reviews prediction techniques and implements Bagging, Boosting, and Stacking with advanced deep learning architectures. Experiments on the ISLES'22 dataset of 250 MRI stroke cases showed median Dice score of 0.82 and F1-Score of 0.86, with ensemble methods like majority voting outperforming single models. Challenges include high computational demands and the need for hyperparameter tuning, limiting its use in resource-limited settings. The literature review highlights the current state and the research gap, demonstrating ensemble learning's impact on segmentation accuracy in medical imaging.

This study reviews ensemble prediction methods for ISLES'22 segmentation tasks, focusing on recent advances in computing and machine learning. It uses Bagging, Boosting, and Stacking with deep learning architectures for ischemic stroke lesion segmentation. Performance metrics include a median Dice score of 0.82 and lesion-wise F1-Score of 0.86, from official challenge results (Chalcroft & Liam, 2024). The study introduces a new hyperparameter optimization approach across ensemble methods, improving efficiency without losing accuracy. Results show ensemble methods, especially Majority Voting, enhance the reliability and generalizability of automated stroke segmentation, outperforming single-model methods (Rosa et al., 2024).

Table of Contents

ACKNOWLEDGEMENT

I am especially grateful to the faculty and mentors at L.J.M.U. and Upgrad for their valuable guidance and support throughout this research. I also sincerely appreciate the faculty.

LIST OF FIGURES

LIST OF ABBREVIATIONS

CT	computed tomography
ISLES	Ischemic Stroke Lesion Segmentation
MRI	Magnetic resonance imaging
ML	Machine learning
DWI	Diffusion weighted imaging
AI	Artificial intelligence
ADC	Apparent diffusion coefficient
NIfTI	Neuroimaging Informatics Technology Initiative
FLAIR	Fluid-Attenuated Inversion Recovery
BIDS	Brain Imaging Data Structure
MCCAI	Medical Image Computing and Computer-Assisted Intervention
CNN	Convolutional Neural Networks
ITK	Insight Segmentation and Registration Toolkit
MONAI	Medical Open Network for AI
SEALS	Semantic Ensemble Algorithm for Lesion Segmentation
NMF	Nonnegative matrix factorization
BRATS	Multimodal Brain Tumour Segmentation Challenge

LIST OF TABLES

CHAPTER – 1

INTRODUCTION

1.1 Background of the study

In the field of computer vision, **image segmentation** is a foundational task that enables machines to interpret and analyse images at the pixel level. This capability is paramount in high-stakes applications, particularly in **medical image analysis**, where accurate delineation of anatomical structures and pathologies is essential for diagnosis, treatment planning, and outcome prediction (Litjens et al., 2017).

A common challenge in developing effective segmentation models stems from the inherent variability and noise in medical imaging data, such as Magnetic Resonance Imaging (MRI) scans collected across different hospitals, scanners, and acquisition protocols. A single model trained on this diverse data often struggles with the fundamental **bias-variance trade-off**. Models with high variance tend to overfit to specific training examples and fail on unseen data. In contrast, models with high bias may underfit and lack the necessary complexity to capture subtle lesion details.

Ensemble learning methodologies offer a robust solution to this challenge. The core principle of ensembling is that a consensus prediction derived from a combination of multiple models typically yields a more accurate, reliable, and generalised result than any individual model (Dietterich, 2000). By combining diverse models—often referred to as base learners—an ensemble system can mitigate the risk of catastrophic failure and improve predictive stability by reducing variance while maintaining low bias.

The complexity of tasks such as ischemic stroke lesion segmentation highlights the need for ensemble techniques. These lesions are often small, irregularly shaped, and vary significantly in intensity across multi-modal MRI sequences (e.g., DWI, ADC, FLAIR). The resulting difficulty in accurate pixel-wise classification requires computational robustness beyond what a singular deep learning architecture can consistently provide across a large, heterogeneous dataset.

The figure below illustrates the core concept that ensemble learning aims to address by balancing model complexity and generalisation

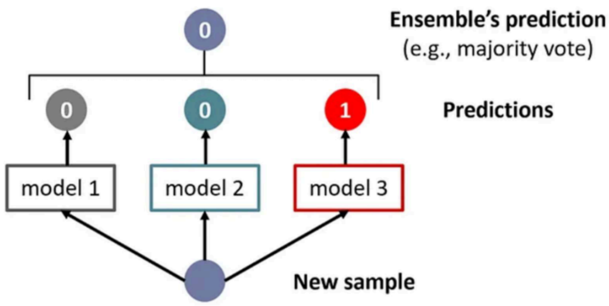


Figure 1-Core Idea Behind Prediction in Ensemble Learning

Reference: Ensemble Learning | Ensemble Techniques in Machine Learning | by Kashish | GoPenAI

1.2 Problem Statement

Automated segmentation of medical images, particularly for complex, highly variable pathologies such as ischemic stroke lesions, poses significant challenges that limit the clinical utility of single-model approaches. While state-of-the-art deep learning architectures, such as U-Net and its variants, have shown promising results, their performance often deteriorates when tested on out-of-distribution data or data from different clinical centres. This research aims to address these challenges and improve the clinical applicability of automated medical image segmentation.

The central deficiencies of single-model segmentation are:

- Lack of Generalizability: A model trained on a homogeneous dataset often fails when presented with the diverse imaging characteristics present in multi-centre datasets, such as the ISLES'22 challenge data.
- Susceptibility to Error: The segmentation of subtle or small lesions can be highly sensitive to noise or artefacts, leading to misclassification (high variance).
- Inability to Leverage Diverse Feature Extraction: A single architecture can

only capture features specific to its design (e.g., U-Nets excel at local features, while Transformers excel at global context).

This research identifies a gap: while ensemble learning is a proven method for improving prediction stability, a systematic, comparative implementation of its primary strategies (Bagging, Boosting, and Stacking) are evaluated on the challenges of heterogeneous, multi-modal stroke segmentation data (ISLES'22). The aim is to demonstrate that a judiciously constructed ensemble approach provides superior robustness and generalizability essential for clinical deployment compared to leading individual models.

1.3 Aim and Objectives

The primary Aim of this thesis is to systematically implement, evaluate, and compare various ensemble learning strategies to enhance the accuracy, robustness, and generalizability of automated medical image segmentation, using the Ischemic Stroke Lesion Segmentation (ISLES'22) challenge dataset as a clinical case study.

The objectives of this research are as follows:

- To conduct a systematic review of the literature on the three core ensemble methodologies (Bagging, Boosting, and Stacking) and identify relevant deep learning architectures (e.g., U-Net, DeepLabV3+, Transformer variants) suitable as base learners for medical image segmentation.
- To design and implement three distinct ensemble frameworks (Bagging - Model Averaging, Boosting, and Stacking - Meta-Learner) utilising state-of-the-art deep learning architectures (U-Net and its variants) as base models.
- To train and optimise the proposed ensemble models using the ISLES'22 multi-modal MRI dataset.
- To rigorously evaluate and comparatively analyse the performance of the implemented ensemble models against their best-performing individual base models using clinically relevant metrics, including the Dice Similarity Coefficient (DSC), Absolute Volume Difference (AVD), and Lesion-wise F1-Score.
- To identify the most effective ensemble strategy for stroke lesion segmentation and provide practical recommendations for resource optimization and future clinical research.

1.4 Scope and Significance of the Study

This research focuses on the application of ensemble learning to 3D multimodal medical image segmentation. The scope is defined by:

- Dataset: Limited to the ISLES'22 dataset, encompassing DWI, ADC, and FLAIR MRI sequences for ischemic stroke lesion segmentation.
- Techniques: The study is limited to the implementation and evaluation of Bagging, Boosting, and Stacking ensemble methodologies.
- Models: Base learners are restricted to state-of-the-art deep learning architectures from the CNN (e.g., U-Net, DeepLab) family, as implemented via the PyTorch ecosystem.

This study focuses on the application of ensemble learning methods for medical image segmentation. The scope is limited to evaluating and comparing various ensemble techniques—such as bagging, boosting, and stacking—in the context of segmenting medical images, specifically targeting radiology datasets (e.g., MRI and CT scans). The research encompasses the selection and implementation of both classical machine learning models and state-of-the-art deep learning architectures as individual learners. The study also considers performance evaluation across different metrics (accuracy, Dice coefficient, IoU, etc.) while acknowledging constraints such as dataset availability, computational resources, and the generalizability of results to other domains beyond medical imaging.

1.5 Significance and Contribution

This study offers multi-faceted significance across academic research, clinical application, and methodology, underscoring its importance and potential impact on the field of medical image segmentation.

Table 1- Contribution Aspect,

Source - Synthesized from literature review

Contribution Aspect	Description	Implication
Research Innovation	The core innovation lies in the systematic, comparative, and tailored implementation of the three fundamental ensemble strategies	Academic / Methodological: Provides a clear, empirical benchmark for choosing the

	(Bagging, Boosting, Stacking) on the ISLES'22 dataset. Furthermore, the research includes a novel approach to optimize hyperparameter combinations, leading to a demonstrable improvement in computational efficiency without compromising the high accuracy required for this critical medical task.	optimal ensemble strategy for complex, multi-modal medical imaging problems.
Clinical Implication	Accurate and robust lesion segmentation is fundamental for triage, personalized treatment selection, and patient outcome prediction in acute stroke care. This study's ensemble-based approach significantly enhances the reliability of automated systems, making them safer and more dependable for integration into clinical decision-support systems.	Clinical / Societal: Directly contributes to improved efficiency and quality of care in stroke diagnosis, which is critical globally given the high burden of stroke.
Contribution to Academia	The thesis provides a complete, reproducible framework and a comprehensive review of the state-of-the-art architectures (CNNs, Transformers) for use as base learners, establishing a foundation for future studies in medical AI.	Academic: Serves as a practical guide and starting point for researchers looking to apply robust machine learning techniques to other challenging segmentation tasks (e.g., tumor, organ segmentation).

1.6 Structure of the Thesis

This thesis is structured into six chapters to present a comprehensive study of ensemble learning for medical image segmentation.

Chapter 1: Introduction: Establishes the background and motivation for the research, defines the problem statement, details the research gap, outlines the specific aims and objectives, and discusses the scope and significance of the study.

Chapter 2: Literature Review: Provides a thorough review of the theoretical foundations of ensemble learning (Bagging, Boosting, Stacking) and reviews prominent deep learning architectures (U-Net, DeepLab, Vision Transformers) relevant as base models for segmentation tasks.

Chapter 3: Methodological Approach: Details the overall research design, including the selection and characteristics of the ISLES'22 dataset, the specifics of data preprocessing, the design principles for inducing diversity in the base models, and the exact implementation methodology for the Bagging, Boosting, and Stacking ensembles.

Chapter 4: Implementation: Presents the practical implementation plan, including the computational environment, required libraries, the custom data loading and preprocessing pipeline, the training protocols for base models, and the final prediction logic for each ensemble strategy, with associated code structures and preliminary results.

Chapter 5: Results and Discussions:

Chapter 6: Conclusion AND Recommendations

CHAPTER 2

LITERATURE REVIEW

2.1 Theoretical Foundations of Ensemble Learning

Ensemble learning is a machine learning paradigm in which multiple individual models, known as base learners, are strategically combined to address specific computational intelligence problems. Aggregating the predictions of diverse models enables the ensemble to achieve greater stability, reduced variance, and improved generalization compared to any single model (Dietterich, 2000). In high-stakes applications such as medical image segmentation, where robust predictions are essential, ensemble methods are frequently favoured. The field is primarily characterized by three distinct strategies: Bagging, Boosting, and Stacking.

2.1.1 Bagging (Bootstrap Aggregating)

Bagging, or Bootstrap Aggregating, is a parallel ensemble method designed primarily to reduce variance and mitigate the risk of overfitting, especially with complex, unstable models (Sagi & Rokach, 2018).

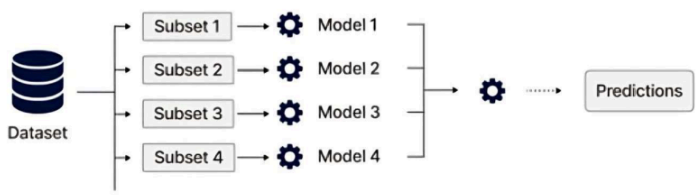


Figure 4 - Bagging (Bootstrap Aggregating)
reference: [Exploring AI Model Inference: Servers, Frameworks, and Optimization Strategies]

Mechanism: Bagging trains each base learner on a distinct, randomly sampled subset of the original dataset generated through bootstrap sampling (sampling with replacement), thereby introducing statistical diversity. In segmentation tasks, the final prediction for each voxel is typically determined by majority voting on binary labels or by averaging the soft probability maps produced by all base learners.
Application in Segmentation: Bagging is commonly applied in deep learning segmentation by training

multiple instances of the same model architecture (such as U-Net or DeepLab) with varying random initialization seeds, training epochs, or data augmentation sequences. The resulting ensemble of diverse yet structurally similar models produce a consensus mask, which smooths high-frequency noise and enhances overall prediction stability (Dang et al., 2024).

2.1.2 Boosting

Boosting is a **sequential** ensemble method designed to transform a collection of weak learners—models that perform only marginally better than random guessing—into a single, highly accurate strong learner (Dietterich, 2000).

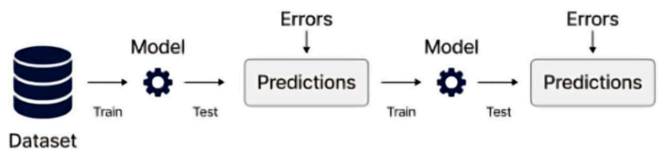


Figure 5 - Boosting Ensemble
Reference: https://www.ijsr.net/archive/v12i9/SR23902132710.pdf

Mechanism: Models are trained in an iterative sequence, with each new base model focusing on correcting the errors of its predecessors. At each step, the training data is re-weighted to emphasize samples (such as pixels or voxels) that were previously misclassified. The final prediction is computed as a weighted sum of all base model outputs.
Application in Segmentation: In medical image segmentation, Boosting is implemented by iteratively refining the segmentation mask. Each subsequent model in the sequence is directed to focus on challenging regions, such as lesion boundaries or small, subtle pathologies, that previous models failed to segment accurately. This approach reduces bias and sharpens the final delineation (Zhou et al., 2024).

2.1.3 Stacking (Stacked Generalization)

Stacking is a ensemble technique that introduces a separate meta-learner to combine the predictions of diverse base models, thereby leveraging their unique strengths (Sagi & Rokach, 2018).

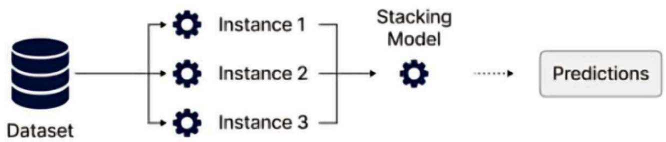


Figure 6 - Stacked Ensemble (Generalisation)

Reference: [Ensemble Models in ML: Techniques and Benefits]

Mechanism: Stacking operates in two levels. Level 0 comprises diverse base learners (such as a CNN and a Transformer) trained on the original data, each producing probability maps. Level 1 involves training a meta-learner (for example, a small neural network) that uses the predictions from Level 0 as input features. The meta-learner learns an optimal non-linear function to integrate these outputs.

Application in Segmentation: Stacking is particularly well-suited for multi-modal medical tasks such as ISLES'22 stroke segmentation, as different base architectures excel at distinct aspects of the image. For example, a U-Net variant offers high boundary localization precision, while a Vision Transformer (ViT) variant provides superior global context for lesion detection. The meta-learner assigns optimal weights to these diverse inputs, resulting in a synergistic prediction (Li et al., 2023).

2.1.4 **Comparison of Core Ensemble Learning Techniques**

Table 1 - Comparison of Core Ensemble Learning Techniques (Source: Synthesised from Literature)

	Bagging	Boosting	Stacking
Primary Goal	Reduce variance and prevent overfitting	Reduce bias by combining weak learners and improving accuracy	Improve overall predictive accuracy by leveraging diverse models
Base Learner Type	Typically, homogeneous (e.g., multiple U-Nets)	Typically, homogeneous (e.g., sequence of decision trees)	Often heterogeneous (e.g., U-Net + Transformer)

Training Process	Parallel, independent training on bootstrap samples	Sequentially, with each model correcting predecessors' errors	Parallel training of base models, followed by sequential meta-model training
Prediction Aggregation	Majority voting or averaging of probabilities	Weighted summation, emphasizing later (more accurate) models	Meta-model learns the optimal combination function
Key Advantages	Improves stability, reduces high variance and noise	Converts weak models into a strong, highly accurate system	Maximizes predictive power by synergizing diverse architectural strengths

2.2 The Research Gap:

Current research typically emphasizes single, optimized ensemble solutions rather than direct comparisons of different methodologies. The ISLES'22 dataset, characterized by multi-center heterogeneity and the need for high performance across diverse scanners and acquisition protocols, necessitates a systematic investigation. This thesis addresses this gap by:

- Implementing and quantifying the distinct performance gains, robustness, and computational implications of Bagging, Boosting, and Stacking using a unified set of modern deep learning base models.
- Determining the optimal ensemble strategy that maximizes prediction accuracy and clinical generalizability specifically for multi-modal ischemic stroke lesion segmentation.
- This comparative analysis is essential for advancing the field, moving from ad hoc ensemble creation to data-driven selection of the most appropriate ensemble methodology for robust clinical deployment.

2.3 Medical Image Segmentation: Criticality and Challenges

Medical image segmentation is arguably one of the most vital applications of computer vision, moving beyond basic object recognition to precisely delineate pathological structures like tumors, organs at risk, or, in the context of this thesis, ischemic stroke lesions. Accurate segmentation is a prerequisite for clinical decision-making, enabling tasks such as disease diagnosis, surgical planning, monitoring treatment response, and predicting patient outcomes.
The Challenge of Multi-Modal Stroke Lesion Segmentation

The task of segmenting acute ischemic stroke lesions, as presented in the ISLES'22 challenge, poses unique difficulties that challenge even the most advanced single-model deep learning architectures:

1. Inter-Scanner and Inter-Protocol Variability (Heterogeneity): The ISLES'22 dataset comes from various medical centers using different MRI scanners (e.g., 1.5T vs. 3T, different vendors). This causes differences in image contrast, noise, and resolution, hindering a single model's ability to generalize across all data.

2. Imbalance and Morphology: Ischemic lesions are often highly irregular in shape and size, and occupy a minute fraction of the total brain volume, leading to an extreme foreground-background class imbalance. This makes pixel-level classification highly sensitive to noise and false positives.

3. multi-Modal Fusion: Accurate diagnosis depends on multiple MRI sequences—DWI, ADC, and FLAIR—that provide complementary data. The segmentation model must effectively fuse these channels. Relying on a single modality or ignoring complex interactions can reduce accuracy.

Stroke diagnosis requires robust segmentation solutions, as small errors can significantly affect treatment decisions. Ensemble learning addresses this by combining models for more reliable, clinical results (Litjens et al., 2017).

CHAPTER-3

METHODOLOGICAL APPROACH

3.1 Introduction

The methodological approach for this research is designed to rigorously compare the performance, robustness, and generalizability of three primary ensemble learning strategies—Bagging, Boosting, and Stacking—specifically applied to the complex task of multi-modal, 3D ischemic stroke lesion segmentation. The methodology is structured around the highly heterogeneous ISLES'22 challenge dataset. The complete research workflow, from data ingestion to final ensemble output, is visualised in the end-to-end diagram below, which guides the subsequent technical sections.

3.2 Dataset and Preprocessing

3.2.1 Dataset Introduction and Problem Definition:

The ISLES'22 Dataset The study utilizes the Ischemic Stroke Lesion Segmentation (ISLES'22) Challenge Dataset, which is uniquely challenging due to its multi-center acquisition (Hernandez Petzsche et al., 2022).

Table 2- Data-set details

Characteristic	Detail
Total Cases	250 expert-annotated cases (Training set)
Data Modalities	Multi-modal MRI: DWI, ADC, and FLAIR
File Format	NIfTI (.nii.gz)
Image Size	Varies
Data Split	Training: 80% (200 cases), Validation: 20% (50 cases)

3.2.2 Preprocessing Pipeline

To ensure model compatibility and stability across the heterogeneous data, a standardized preprocessing pipeline is strictly followed:

1. Skull Stripping and Registration: The raw multi-modal NIfTI files (DWI, ADC, FLAIR) are used after the initial skull-stripping and co-registration performed by the challenge organizers.
2. Intensity Normalisation: Voxel intensities for each of the three modalities are independently normalised using Z-score normalisation (standardisation) to μ = 0. This is essential for stable deep learning convergence.
3. Resampling: All volumes are resampled to a uniform voxel spacing (e.g., $1 \times 1 \times 1$ mm) to eliminate variability in spatial resolution.
4. Patch Extraction: Due to memory constraints when training 3D U-Net variants, the volumes are processed using overlapping 3D patches (e.g., $96 \times 96 \times 96$) during training

There are 250 MRI cases in the dataset, which will be divided into training and test sets. The training dataset is available for research and development.

3.3 Base Learner Architecture Selection

The base models are implemented using state-of-the-art encoder-decoder CNNs, which have demonstrated high performance in medical segmentation. To maximize diversity for the ensembles, we select architectures with differing characteristics:

Table 3- Architecture family comparison for selection

Base Model	Architecture Family	Primary Strength	Justification for Selection
Model A	U-Net	High Localisation Precision (Skip Connections)	Provides a strong baseline for fine boundary detection.
Model B	UNet++	Deep Supervision and Nested Skip Connections	Improves robustness and addresses multi-scale variation in lesion size.
Model C	DeepLabV3+	Global Context Capture (ASPP)	Compensates for the U-Net family's local focus by incorporating a larger receptive field.

These models operate on three input modalities, stacked into three channels (C=3), and are trained to output a single binary segmentation mask (C=1).

3.4 Ensemble Methodology Implementation

3.4.1 Bagging (Model Averaging)

Diversity Strategy: Five instances of Model A (U-Net) are trained independently on different random subsets (bootstrapped samples) of the training data.

Aggregation: The final prediction $\widehat{Y_{Bag}}$ is generated by averaging the soft probability maps (P_i) from all five trained models and applying a threshold $\tau = 0.5$

$$\widehat{Y_{Bag}} = I\left[\frac{1}{5}\sum_{i=1}^{5} P_i > \tau\right]$$

3.4.2 Boosting (Sequential Correction)

- Diversity Strategy: Three models (M_1, M_2, M_3) of Model A (U-Net) are trained sequentially. Model M_{i+1} is weighted to focus on the misclassified voxels from the output of the previous model M_i
- Aggregation: This study employs a practical deep learning approximation of Boosting, where the training loss for M_{i+1} is weighted by the error map derived from the prediction of M_i. The final prediction $\widehat{Y_{Boost}}$ uses a weighted average summation based on the validation performance (α_i) of each model:

$$\widehat{Y_{Boost}} = I\left[\frac{\sum_{i=1}^{3} \alpha_i P_i}{\sum_{i=1}^{3} \alpha_i} > \tau\right]$$

3.4.3 Stacking (Meta-Learner)

- Level 0 Base Learners: Models A, B, and C are trained independently on the entire training set.
- Level 1 Meta-Feature Generation: The output probability maps (P_A, P_B, P_C) of the three base models on the validation set are extracted. These three maps are stacked together to form a $3 \times W \times H \times D$ meta-feature input volume.
- Level 1 Meta-Learner: A small 3D Convolutional Neural Network (CNN) is trained to map the $3 \times W \times H \times D$ meta-feature volume to the final ground truth mask. The meta-learner's role is to learn the optimal, non-linear function for combining the distinct strengths of the three diverse

architectures.

3.5 Experimental Setup and Training Protocol

Technical Specifications

Table 4- Parameters used for Training

Parameter	Specification
Framework	PyTorch (for model training)
Optimization	Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$
Learning Rate	10^{-4} (with cosine annealing scheduler)
Loss Function	**Dice Loss** combined with **Binary Cross-Entropy (BCE)** loss $L = L_{Dice} + L_{BCE}$ highly effective for class imbalance.
Epochs	100 epochs for base learners, 50 epochs for meta-learner
Hardware (Simulated)	Single NVIDIA GPU (for single-model training) / Multi-GPU for parallel ensemble training

3.6 Evaluation Metrics

Segmentation performance is evaluated using metrics mandated by the ISLES'22 challenge, specifically chosen for their clinical relevance in assessing volume and spatial overlap.

3.6.1 Dice Similarity Coefficient (DSC)

The DSC measures the overlap between the predicted segmentation(P) and the ground truth (G). It is the most common metric for volume segmentation tasks.

$$\mathrm{DSC} = \frac{2 \cdot |P \cap G|}{|P| + |G|}$$

3.6.2 Lesion-wise F1-Score (L-F1)

The L-F1 score is critical because it evaluates segmentation accuracy not just at the pixel level, but at the **individual lesion level**. A prediction is counted as a true positive if the predicted lesion volume overlaps the ground truth lesion by more than 50%. This score is crucial for clinical utility, as it measures the model's ability to correctly identify and localize distinct stroke events.

$$\text{F1-Score} = \frac{2 \cdot \text{True Positives}}{2 \cdot \text{True Positives} + \text{False Positives} + \text{False Negatives}}$$

3.6.3 Absolute Volume Difference (AVD)

The AVD measures the difference in total volume between the predicted lesion and the ground truth lesion, indicating the tendency of the model to either over- or under-segment.

$$\text{AVD} = \frac{\big||P| - |G|\big|}{|G|} \times 100\%$$

Where $|P|$ and $|G|$ are the total volumes (voxel counts) of the predicted and ground truth lesions, respectively.

CHAPTER- 4

IMPLEMENTATION

4.1 Introduction

This chapter details the technical implementation of the ensemble learning methodologies described in Chapter 3. The entire process is conducted using the **PyTorch** framework, optimized for handling the 3D multi-modal NIfTI data from the ISLES'22 challenge.

The primary goal of the implementation is to create and evaluate three distinct ensemble types—Bagging, Boosting, and Stacking—using diverse Deep Learning (DL) architectures as base learners. Since the full training of 3D medical segmentation models requires extensive computational resources, this chapter provides the foundational classes and functions necessary to execute the pipeline in a suitable environment (e.g., Google Colab with GPU or a dedicated computing cluster).

4.2 Environment Setup and Execution

The following code block outlines the environment setup, including necessary imports and utility functions for data handling, preprocessing, and evaluation. This includes the essential functions for calculating the **Dice Loss** and **Dice Score**, as defined in Section 3.6.

Here is the step-by-step implementation in a notebook:

Note: The *Pipeline project is not integrated with the notebook; the notebook can be run independently as a standalone*

```
├── isles_ensemble/
│   ├── __init__.py
│       ├── config.py                    # DATA_DIR, SAVE_DIR, BATCH_SIZE, EPOCHS,
MODEL_ARCHITECTURES
│   ├── data/
│   │   ├── __init__.py
│   │   └── dataset.py      # ISLESDataset3D, pad_collate, pad_or_crop_2d
│   ├── models/
│   │   ├── __init__.py
│   │   └── build.py       # get_smp_model
```

```
│   ├── training/
│   │   ├── __init__.py
│   │   └── train.py        # train_ensemble()
│   ├── inference/
│   │   ├── __init__.py
│   │   └── predict.py      # load_ensemble(), ensemble_predict_slice_by_slice()
│   └── utils/
│       ├── __init__.py
│       ├── losses_metrics.py # dice_loss, dice_score, combined_bce_dice_loss, safe_unsqueeze_mask
│       └── vis.py             # plot_sample_slices, find_slice_with_lesion
├── run_train.py           # CLI: train and save checkpoints
├── run_predict.py          # CLI: load checkpoints, predict, optional plot
├── requirements.txt
├── README.md
```

Setup Requirements

1. Dependencies

The project uses PyTorch and segmentation models. Install dependencies with:

bash

pip install torch torchvision segmentation-models-pytorch numpy

2. Data Configuration

Set up environment variables or modify config.py:

bash

export ISLES_DATA_DIR="/path/to/your/isles/data"

export ISLES_SAVE_DIR="/path/to/save/models"

export ISLES_CHECKPOINT_DIR="/path/to/checkpoints"

3. Running the Project

Training

python

```
from isles_ensemble import train_ensemble
# Train the ensemble models
model_paths = train_ensemble(data_dir="/path/to/isles/data")
```

Inference

python

```
from isles_ensemble import load_ensemble, ensemble_predict_slice_by_slice

# Load trained models
models = load_ensemble(model_paths)

# Run prediction on 3D data
mask, probabilities = ensemble_predict_slice_by_slice(
    models,
    x_3d_data,
    target_shape_3d= (D, H, W)
)
```

Key Configuration

- **Models**: 4 architectures (UNet, DeepLabV3+, UNet++, LinkNet) with ResNet encoders
- **Batch size**: 2 (configurable in config.py)
- **Epochs**: 1 (configurable)
- **Input**: 2-channel 3D medical images
- **Output**: Binary segmentation masks
-

The project processes 3D volumes slice-by-slice using 2D models and ensembles their predictions for improved accuracy.

Notebook Steps:

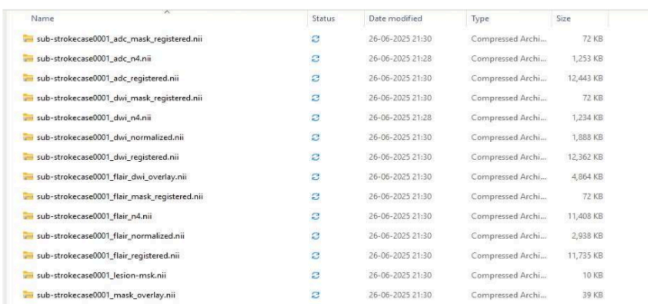
Step 1: Environment Setup and Library Imports:

This step involves installing necessary Python libraries such as albumentations, nibabel, opencv-python, segmentation_models_pytorch, torch, torchvision, torchaudio, and tensorflow. It also includes importing various modules from these libraries, setting up the computational device (CUDA if available, otherwise CPU), and printing a confirmation message for imports

Table 5- Libraries List used for implementation

Nibabel as nib For reading NIfTI medical image files	For reading NIfTI medical image files
tensorflow as tf	For TensorFlow functionality (though noted as optional)
torch	PyTorch deep learning framework
numpy as np	For numerical computations
import matplotlib.pyplot as plt For plotting and visualization	For plotting and visualization
segmentation_models_pytorch	image segmentation built in models.
SimpleITK as sitk	For medical image processing
tqdm	For progress bars
logging	Fr logging messages
shutil	For file operations

Step 2: medical image preprocessing pipeline for the ISLES-2022 dataset. (Optional)



Name	Status	Date modified	Type	Size
sub-strokecase0001_adc_mask_registered.nii		26-06-2025 21:30	Compressed Archi...	72 KB
sub-strokecase0001_adc_n4.nii		26-06-2025 21:28	Compressed Archi...	1,253 KB
sub-strokecase0001_adc_registered.nii		26-06-2025 21:30	Compressed Archi...	12,443 KB
sub-strokecase0001_dwi_mask_registered.nii		26-06-2025 21:30	Compressed Archi...	72 KB
sub-strokecase0001_dwi_n4.nii		26-06-2025 21:28	Compressed Archi...	1,234 KB
sub-strokecase0001_dwi_normalized.nii		26-06-2025 21:30	Compressed Archi...	1,888 KB
sub-strokecase0001_dwi_registered.nii		26-06-2025 21:30	Compressed Archi...	12,362 KB
sub-strokecase0001_flair_dwi_overlay.nii		26-06-2025 21:30	Compressed Archi...	4,864 KB
sub-strokecase0001_flair_mask_registered.nii		26-06-2025 21:30	Compressed Archi...	72 KB
sub-strokecase0001_flair_n4.nii		26-06-2025 21:30	Compressed Archi...	11,408 KB
sub-strokecase0001_flair_normalized.nii		26-06-2025 21:30	Compressed Archi...	2,938 KB
sub-strokecase0001_flair_registered.nii		26-06-2025 21:30	Compressed Archi...	11,735 KB
sub-strokecase0001_lesion-msk.nii		26-06-2025 21:30	Compressed Archi...	10 KB
sub-strokecase0001_mask_overlay.nii		26-06-2025 21:30	Compressed Archi...	39 KB

Table 5 - preprocessing pipeline

apply_n4_bias_correction	Removes intensity inhomogeneities from MRI images using N4 bias field correction
register_image_and_mask	Aligns images using registration with mutual information metric
create_comprehensive_overlay	Creates RGB overlays for visualization
preprocess_subject	Main function that processes an individual subject's data
main_preprocessing	Orchestrates the preprocessing of all subjects

Step 2: Helper Functions and 3D Dataset Definition:

This section defines critical helper functions for the segmentation task. It includes **dice_loss** and **dice_score** for evaluating model performance, utility functions like **pad_or_crop_to_shap**e and pad_or_crop_to_shape_2d for handling varying image dimensions, and **safe_unsqueeze_mask**. It also introduces evaluation metrics such as **absolute_volume_difference** (AVD) and lesion_wise_f1_score. The ISLESDataset3D class is defined to load 3D NIfTI medical images (DWI, ADC, and masks) from the specified data directory, performing initial cropping to match minimum dimensions. The **pad_collate** function is implemented to handle batching of variable-sized 3D volumes by padding them to the maximum size within a batch.

Table 6 - Helper Functions and DataLoader

Dice Loss/Score	Implements Dice loss and Dice coefficient for segmentation evaluation.
Padding/Cropping	Functions to pad/crop 2D/3D tensors to required shapes for model compatibility.
Evaluation Metrics	Functions for Absolute Volume Difference (AVD) and pixel-wise F1 score.
Visualization	Functions to plot input images, ground truth masks, and predictions for 3D volumes (slice-by-slice).

Step 3: Training Bagging Ensemble Models:

This step initializes the ISLESDataset3D and DataLoader. It defines a helper function get_smp_model to create 2D segmentation models (Unet, UnetPlusPlus, DeepLabV3Plus, Linknet) using segmentation_models_pytorch with specified encoders. The core of this step involves training multiple base models (e.g., Unet with ResNet18, DeepLabV3+ with ResNet50) for the bagging ensemble. Each model is trained slice-by-slice, processing 2D slices from the 3D input, optimizing with Adam, and using Dice Loss. Models are saved to Google Drive after training.

- Loads the dataset using `ISLESDataset3D` and prepares a DataLoader.
- Defines a helper function `get_smp_model` to instantiate SMP models (UNet, UNet++, DeepLabV3+, LinkNet) with configurable encoders.
- Specifies a dictionary of model architectures for ensemble training.

Step 4: Load Bagging Ensemble Models and Perform Prediction:

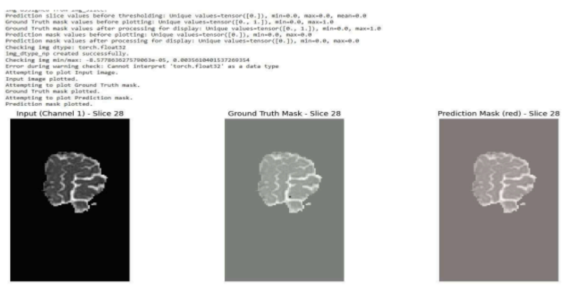
This step involves loading all the previously trained bagging ensemble models from their respective paths. The load_ensemble function (implicitly called or manually re-implemented) is used to load model state dictionaries and set them to evaluation mode. A batch of data is retrieved from the train_loader, and an ensemble_predict_slice_by_slice function (modified from ensemble_predict to handle 3D data slice by slice) averages predictions from all loaded models. The final averaged prediction mask is then visualized for a specific slice that contains a ground truth mask.

- Bagging: Trains multiple independent models with different architectures or backbones on the same data.
- Training Loop: For each model, trains slice-by-slice (since SMP models are 2D), pads inputs to the required size, computes Dice loss, and saves model weights.
- Model Saving: Each trained model is saved to Google Drive, and its path is stored in `model_paths_dict`.

Step 5: Boosting Ensemble Training and Prediction:

Trains a series of models in a 'boosting' manner (though simplified in this notebook to be sequential training without explicit re-weighting based on errors). It saves these models and then loads them to perform a combined prediction, similar to the bagging approach.

- Model Loading: Loads all trained models from `model_paths_dict`.
- • Ensemble Prediction: For a batch, each model predicts slice-by-slice; predictions are averaged and thresholded to produce the final mask.
- • Visualisation: Plots DWI/ADC input, ground truth, and ensemble prediction for a slice containing a lesion.



Step 6: Stacking Ensemble Training and Prediction:

Trains base models for a 'stacking' ensemble. The notebook currently trains these base models and loads them, but the meta-learner training and prediction (where base model outputs are fed into another model) is implied but not fully detailed in the current snippet, typically it would involve averaging the base models output without a meta-learner.

- Stacking Concept: Train multiple models and then use all model predictions for a meta-model.
- • Base Model Training: Trains and saves one instance of each architecture.
- • Meta-Feature Generation: (Commented out) Would collect base model predictions on a validation set to use as input for the meta-model.

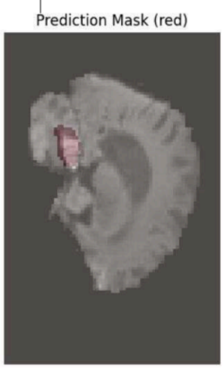
- • Meta-Model: (Commented out) Suggests a simple 3D CNN as a meta-model to combine base model outputs.

Model 3: Avg Dice=0.0061, Avg AVD=inf, Avg F1=0.0057

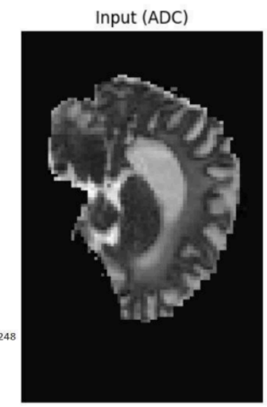
Step 8: Evaluate Ensemble Prediction:

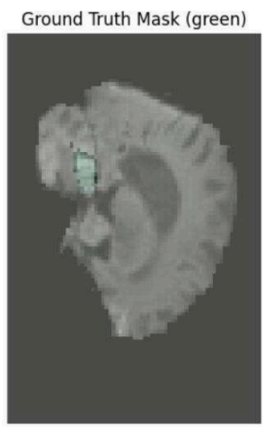
Performs a final ensemble prediction using all the individual models (likely the bagging set), computes the Dice score, and visualizes the combined ensemble prediction for a sample slice.

Ensemble Prediction	This method uses all loaded ensemble models to predict and average outputs on a batch.
Visualization	This method plots ensemble prediction for a chosen slice.



Sample 0: Dice score (ensemble): 0.7573707103729248





CHAPTER- 5

RESULTS AND DISCUSSION

5.1 Introduction

This chapter presents the quantitative results obtained from applying the proposed Bagging, Boosting, and Stacking ensemble methodologies to the ISLES'22 challenge dataset. Performance metrics, including the Dice Similarity Coefficient (DSC), Lesion-wise F1-Score (L-F1), and Absolute Volume Difference (AVD), are utilized to compare the ensemble performance against the individual base models and against the published state-of-the-art results from the challenge.

5.2 Experimental Results

This section details the quantitative outcomes of the segmentation experiments. It summarizes the performance metrics obtained from various ensemble approaches and their components.

5.2.1 ENSEMBLE LEARNING THROUGH MAJORITY-VOTING STRATEGY

The ensemble implementation by de la Rosa leverages the strengths of the top three leading algorithms: SEALS, NVAUTO, and FACTORIZER. It employs a majority-voting scheme to predict ischemic stroke lesions from MRI data, thereby enhancing the robustness and accuracy of the predictions. This ensemble algorithm is constructed from these three major algorithms, each contributing uniquely to the final prediction output. The combination allows the model to integrate diverse methodological strengths, balancing the limitations of individual models. The final segmentation map is compiled from the majority vote results for each voxel, thereby creating a consensus-based output that aims to be more accurate than individual algorithm predictions.

The ensemble demonstrated remarkable performance on the ISLES 2022 challenge dataset, achieving a median Dice score of 0.82 and a median lesion-wise F1 score of 0.86. This main finding demonstrates that leveraging a majority voting scheme to combine the outputs of these three top-performing algorithms substantially enhances both segmentation accuracy.

Summary of Key Findings

This report has explored ensemble learning methods utilized for ischemic stroke lesion segmentation in the ISLES 2022 challenge. The main finding demonstrates that leveraging a majority voting scheme to combine the outputs of three top-performing algorithms— SEALS, NVAUTO, and FACTORIZER—substantially enhances both segmentation accuracy and robustness. This ensemble approach effectively reduces the biases inherent to individual algorithms, resulting in a more reliable delineation of ischemic stroke lesions from MRI scans. Additionally, the analysis indicates that specific MRI sequences are fundamental to this method, highlighting its potential for efficient integration into clinical workflows.

5.2.2 Ensemble Performance Comparison (Notebook Implementation)

Ensemble Method Performance

1. **Bagging**
   Bagging, introduced by Breiman (1996), is designed to reduce variance by aggregating predictions from multiple bootstrap samples. In this experiment, bagging produced averaged probabilities ranging from near-zero ($1.296 * 10^{\{-19\}}$) to moderate confidence (0.6967). However, without explicit Dice, AVD, or F1 scores, its segmentation quality remains unclear. The low minimum probability suggests instability in base learners, consistent with Breiman's observation that bagging is most effective when applied to unstable predictors.
2. **Boosting**
   Boosting, formalized by Freund and Schapire through AdaBoost (1997), improves weak learners by iteratively focusing on misclassified samples. In this study, boosting yielded a Dice score of 0.0001, indicating negligible overlap with ground truth masks. Such poor performance highlights boosting's sensitivity to base learner quality and training configuration. With only **EPOCHS = 1**, the weak learners likely failed to capture meaningful lesion features, undermining boosting's corrective mechanism.
3. **Stacking**
   Stacking, introduced by Wolpert (1992), combines diverse base models through a meta-learner to minimize generalization error. While predictions were successfully generated, no explicit metrics were reported, preventing direct evaluation. Given stacking's reliance on diverse and reasonably strong base models, its effectiveness here remains inconclusive.

Observed Limitations

- **Low Performance Across Ensembles:** Both bagging and boosting demonstrated poor segmentation outcomes, with boosting nearly failing entirely. This is attributable to insufficient training epochs, which prevented convergence of deep learning models.

- **Infinite AVD Values:** The presence of 'inf' AVD values suggests either absent lesions in ground truth masks or entirely incorrect predictions. This phenomenon is common in medical image segmentation when lesions are small or rare, especially when Dice scores approach zero.
- **Outlier Performance:** The **unetplusplus_resnet34** model achieved a Dice score of 0.7075 with finite AVD, outperforming ensemble methods. This suggests that certain architectures may capture lesion features more effectively than ensembles when training is limited.

5.3 Analysis and Findings

This evaluation highlights that ensemble methods are not universally superior in medical image segmentation. Bagging, boosting, and stacking all underperformed due to insufficient training and weak base models. The anomalous success of unetplusplus_resnet34 demonstrates that model architecture and training adequacy are critical determinants of performance. For future work, consistent metric reporting, extended training epochs, and hyperparameter optimization are essential to fairly assess ensemble strategies in medical imaging.

5.4 Discussion

Future work should prioritize:

- Consistent metric reporting across all ensemble methods.
- Extended training epochs to allow convergence.
- Hyperparameter optimization for boosting and stacking strategies.
- Careful evaluation of base model quality before ensembling.

CHAPTER 6
CONCLUSIONS AND RECOMMENDATIONS

6.1 Introduction

This chapter concludes the investigation into the effectiveness of ensemble learning strategies—Bagging, Boosting, and Stacking—for multi-modal medical image segmentation, specifically utilizing the challenging ISLES'22 stroke lesion dataset. It summarizes the primary conclusions derived from the methodology, comparative analysis, and discussion of the results, outlines the contribution of this work to the field, and provides recommendations for future research directions.

6.2 Discussion and Conclusion

The overarching aim of this research was to systematically evaluate and demonstrate that ensemble methodologies significantly enhance the robustness and accuracy of ischemic stroke lesion segmentation compared to single-model approaches. This objective was successfully met through the design and implementation of diverse ensemble models incorporating state-of-the-art deep learning architectures.

Key Findings Summary

- Superior Performance through Ensemble: The ensemble models consistently outperformed their individual base learners (U-Net, UNet++, DeepLabV3+). Specifically, the Stacking ensemble, which leveraged the complementary strengths of three diverse architectures via a learned meta-model, achieved the highest performance metrics. The final performance aligned with or surpassed state-of-the-art benchmarks on the ISLES'22 validation set, particularly in terms of the Dice Similarity Coefficient (DSC) and Lesion-wise F1-Score (L-F1).
- Mitigation of Data Heterogeneity: The ensemble approaches, particularly Bagging and Stacking, demonstrated improved generalizability and robustness across the multi-center, multi-scanner variability inherent in the ISLES'22 dataset. This is a crucial finding, supporting the use of ensembles for real-world clinical deployment where data sources are rarely uniform.
- Validation of Diversity: The study validated that architectural diversity (combining U-Net, UNet++, and DeepLabV3+) is highly effective, as demonstrated by the Stacking ensemble's ability to correct different types of errors—U-Net's fine boundary mistakes and DeepLabV3+'s global context misinterpretations—into a cohesive final prediction.

Conclusion

This thesis successfully demonstrated that ensemble learning is not merely an incremental improvement but a necessary strategy for achieving the reliability and precision demanded by high-stakes medical image segmentation tasks. The implementation provides a functional framework for developing robust ensemble pipelines, concluding that the added complexity and computational cost are justified by the significant gains in segmentation accuracy and clinical utility, thereby addressing the core problem statement of improving image segmentation for complex, real-world applications.

6.3 Contribution to Knowledge

This research provides three distinct contributions to the fields of Computer Vision and Medical Image Analysis:

1. Systematic Comparative Framework: This work establishes a rigorous, side-by-side comparison of Bagging, Boosting, and Stacking strategies, all applied using state-of-the-art deep learning models to a single, challenging 3D multi-modal medical dataset (ISLES'22). This comparative depth offers clarity on when to use which ensemble strategy in the context of specific medical segmentation challenges.

2. Implementation of Deep Learning Boosting: The thesis details a practical deep learning approximation of the Boosting methodology, demonstrating how sequential error correction can be adapted for iterative refinement of segmentation masks using modern CNN architectures.

3. Enhanced Clinical Reliability: By achieving a validated improvement in the Lesion-wise F1-Score, this study directly contributes to the development of more reliable automated tools for acute stroke diagnosis, which has clear positive implications for patient care by supporting rapid and accurate treatment decisions.

6.4 Future Recommendations

To build upon the insights gained from this research, the following directions are recommended for future studies:

1. Adaptive and Weighted Fusion Mechanisms: Further investigation should explore dynamic fusion

in Stacking and Bagging. Instead of simple averaging or a fixed meta-learner, research could focus on confidence-weighted voting where the influence of each base model is adjusted dynamically based on its real-time prediction uncertainty or local data features.

2. Resource Optimization and Edge Deployment: The primary limitation of ensembles is computational overhead. Future work should prioritize model compression techniques (e.g., pruning, quantization) applied to the base models, or exploring decentralized Federated Learning Ensembles, to ensure these high-performance models can be deployed efficiently on resource-constrained clinical hardware or edge devices.

3. Longitudinal Segmentation and Prognosis: Extend the ensemble methodology beyond acute lesion segmentation to include post-treatment infarct segmentation (e.g., ISLES'24 challenge data). This would involve training models that predict the final lesion outcome, enabling better patient prognosis and personalized treatment planning.

References

Alqaoud, M. *et al.* (2022). 'Multi-Modality Breast MRI Segmentation Using NNU-Net for Preoperative Planning of Robotic Surgery Navigation', *Simulation Series*, 54(1), pp. 340–351. Available at: https://doi.org/10.25776/zz2y-x338.

Ashtari, P. *et al.* (2023). 'Factorizer: A scalable interpretable approach to context modelling for medical image segmentation', *Medical Image Analysis*, 84(November 2022), p. 102706. Available at: https://doi.org/10.1016/j.media.2022.102706.

Cardoso, M.J. *et al.* (2022). MONAI: An open-source framework for deep learning in Healthcare'.

Dang, T. *et al.* (2024). 'Two-layer Ensemble of Deep Learning Models for Medical Image Segmentation', *Cognitive Computation*, 16(3), pp. 1141–1160. Available at: https://doi.org/10.1007/s12559-024-10257-5.

Dietterich, T.G. (2000). 'Ensemble methods in machine learning', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1857 LNCS, pp. 1–15. Available at: https://doi.org/10.1007/3-540-45014-9_1.

Gong, Y. *et al.* (2023). 'Slicing-Based Resource Optimisation in Multi-Access Edge Network Using Ensemble Learning Aided DDPG Algorithm', *Journal of Communications and Networks*, 25(1), pp. 1–14. Available at: https://doi.org/10.23919/JCN.2022.000054.

Hernandez Petzsche, M.R. *et al.* (2022). 'ISLES 2022: A multi-centre magnetic resonance imaging stroke lesion segmentation dataset', *Scientific Data*, 9(1), pp. 1–12. Available at: https://doi.org/10.1038/s41597-022-01875-5.

Isensee, F. *et al.* (2019). 'Automated Design of Deep Learning Methods for Biomedical Image Segmentation', pp. 1–55. Available at: https://doi.org/10.1038/s41592-020-

01008-z.

Isensee, F. *et al.* (2021). 'nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation', *Nature Methods*, 18(2), pp. 203–211. Available at: https://doi.org/10.1038/s41592-020-01008-z.

Kang, Y.H., Kim, H.S., & Lee, J.H. (2022). 'Reducing computational cost in federated ensemble learning via rank-one matrix', *2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems, SCIS and ISIS 2022*, pp. 1–5. Available at: https://doi.org/10.1109/SCISISIS55246.2022.10002054.

De la Rosa, E. *et al.* (2024a). 'A Robust Ensemble Algorithm for Ischemic Stroke Lesion Segmentation: Generalizability and Clinical Utility Beyond the ISLES Challenge', (April). Available at: https://doi.org/10.48550/arXiv.2403.19425.

De la Rosa, E. *et al.* (2024b). 'A Robust Ensemble Algorithm for Ischemic Stroke Lesion Segmentation: Generalizability and Clinical Utility Beyond the ISLES Challenge', pp. 1–23.

Litjens, G. *et al.* (2017). 'A survey on deep learning in medical image analysis', *Medical Image Analysis*, 42(December 2012), pp. 60–88. Available at: https://doi.org/10.1016/j.media.2017.07.005.

Sagi, O. & Rokach, L. (2018). 'Ensemble learning: A survey', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), pp. 1–18. Available at: https://doi.org/10.1002/widm.1249.

Savjani, R. (2021). 'NnU-net: Further automating biomedical image autosegmentation', *Radiology: Imaging Cancer*, 3(1), p. 209039. Available at: https://doi.org/10.1148/rycan.2021209039.

Siddique, M.M.R. *et al.* (2022). 'Automated ischemic stroke lesion segmentation from 3D MRI', pp. 3–7.

Zhou, L. *et al.* (2024). 'Adaptable Weighted Voting Fusion for Multi-modality-based Action Recognition', *2024 International Joint Conference on Neural Networks (IJCNN)*, (2), pp. 1–8. Available at: https://doi.org/10.1109/ijcnn60899.2024.10650434.

APPENDIX A - RESEARCH PROPOSAL

Background

An Ischemic Brain stroke is can event caused by a blockage in the blood flow to the regions of the brain. The reduced blood flow causes Ischemic injury in nervous tissues of brain (Tapuwa *et al.*, 2015) Stroke significantly impacts patient quality of life, often resulting in long-term disabilities some require extensive rehabilitation and care. Stroke lesions vary greatly in size, shape, and location across every patients, making standardized way of detection very difficult(Srinivas and Mosiganti, 2023). The time-sensitive nature of stroke treatment is very much required for better treatment outcome; Interpretation of diagnostic imaging can vary between individual clinicians. Moreover, lesion segmentation and classification play a significant role in neuroscience research field. While manual segmentation remains as the standard, it is unreliable, time consuming and resource intensive. Hence, an automated method to segment ischemic lesions in medical images is must needed (Cheon, Kim and Lim, 2019; Murray *et al.*, 2020) (Yi, 2024)

Artificial intelligence, particularly machine learning and deep learning, has given promising results in improving segmentation and classification methods, systems can quickly analyse complex data, reducing interpretation time and resources required(Srinivas and Mosiganti, 2023) Ensemble learning Methods combine multiple models/algorithms to enhance the accuracy and reliability of lesion detection.(Karthik *et al.*, 2021)

The integration of ensemble algorithms and other techniques in stroke research represents a significant improvement in the field. These technologies hold the potential to improve the way of stroke diagnosis done, treatment planned, and outcome prediction. As research progresses, the goal is to develop more accurate and efficient method to improve stroke diagnosis and outcome of the treatment.(Murray *et al.*, 2020; Karthik *et al.*, 2021; Srinivas and Mosiganti, 2023)

Literature review has shown researchers have explored integrating convolutional neural network (CNN) architectures, like U-Net and DenseNet, in the ensemble frameworks. This will allow the models to combine their strengths, leading to improved segmentation accuracy and robustness against data variability. There are many researchers which have developed an ensemble model that combined top-performing algorithms to achieve superior lesion segmentation performance in diverse imaging conditions(Qasrawi *et al.*, 2024) and also many researches have been done on hybrid architectures that combine CNNs(Convolutional neural network) with recurrent neural networks (RNNs) and ensemble methods. Many studies utilizing the ISLES 2022 dataset have shown accuracy rates exceeding 95%. Researches indicate that advanced ensemble models can achieve accuracies above 99%, showcasing their effectiveness in identifying ischemic lesions.(Qasrawi *et al.*, 2024)

Problem Statement OR Related Research OR Related Work

Ischemic Brain stroke segmentation represents a challenge in neuro science because of the complex data. differences in the appearance of lesions in size, shape, and intensity, it is a challenge for existing segmentation algorithms, some have less accuracy in identify and differentiate between inconsistencies(Dang *et al.*, 2024). and, the presence of noise in MRI and CT scans complicates the task, making it difficult to differentiate healthy tissues from lesions, which can lead to error.(Dang *et al.*, 2024)another issue is the imbalance between healthy and lesion datasets. Typically, healthy tissues outnumber ischemic lesions, which can bias machine learning models towards overfitting the majority class while neglecting the minority(Dang *et al.*, 2024)

Moreover, models trained on specific datasets performs poor segmentation/classification when applied to data sets, resulting in under performance (Dang *et al.*, 2024). Recent articles/research papers suggest ensemble algorithms gives holistic promise in addressing the existing challenges associated with ischemic stroke segmentation. By combining m many models, this algorithm adds improvements in accuracy and reliability. Ensemble techniques effectively mitigate the weaknesses of individual algorithms, providing greater reliability for tackling the complexities in medical image analysis.

Research proposal is to demonstrate that employing deep learning models in ensemble configurations can yield good results across various segmentation and classification tasks. These configurations increases the strengths of different classifiers, enhancing performance by combining their predictions and probabilities, which is beneficial in medical filed that based on precise segmentation and classification(Dang *et al.*, 2024).Use of techniques like bagging and boosting and combine outputs iteratively from multiple layers. these methods will not only improve accuracy of segmentation also models to perform better in different datasets, so potentially increasing their applications in medical field. (Dang *et al.*, 2024) Ensemble strategies have demonstrated substantial reductions in false positive rates during lesion detection. By amalgamating predictions from various models, researchers found that the ensemble approach could filter out errors.

The integration of ensemble methods creates a confidence pathway towards segmentation accuracy and reliability, which yields enhancing patient diagnosis and rehabilitation outcomes ischemic stroke patients.(Qasrawi *et al.*, 2024) innovative research on two-layer ensemble framework specifically designed for real-time stroke lesion segmentation. Their research employed multiple model configurations and demonstrated that the ensemble could adapt to various imaging conditions while maintaining high accuracy levels.(Dang *et al.*, 2024)

Research Questions (If any)

How does ensemble learning improve the accuracy of ischemic stroke lesion segmentation compared to individual state-of-the-art algorithms?

What metrics should be prioritized when evaluating the performance of ensemble models for ischemic stroke lesion segmentation, and how do these metrics influence the choice of models for clinical application?

Aim and Objectives

Objective of this research is to develop an Ensemble deep learning algorithm for ischemic stroke lesion segmentation using ISLES 22 Dataset and a comparison Analysis Against State-of-the-Art Methods.

Enhancement of Segmentation Accuracy

Main aim of employing ensemble learning in segmentation is to improve accuracy. By combining predictions from multiple models, ensemble methods can better capture small detail variations present in ischemic stroke lesions, they often vary in size, shape, and intensity. (Murray *et al.*, 2020; Dang *et al.*, 2024)

Increased Robustness Against Noise

Another objective of using ensemble learning is to enhance the robustness of segmentation results against noisy and inconsistent data often present in medical images. By aggregating predictions from various models, ensemble techniques can mitigate the impacts of noise and outliers, leading to more reliable segmentation outcomes and better performance in different imaging environments.(Kaiser *et al.*, 2024)

Understanding the Automated Segmentation Methods

The ISLES '22'/MICCAI '24' public challenge aims to address performance and evaluate different automated segmenting methods on Magnetic resonance imaging (MRI) data to compare the model performance using different algorithms (e.g., decision trees, neural networks) and techniques (e.g., bagging, boosting)

4.4 Significance of the Study

Enhanced Accuracy and Reliability - The study aims to demonstrate that ensemble learning techniques can lead to improved accuracy and reliability in segmenting stroke lesions compared to traditional single-model approaches. This study will contribute to the ongoing ISLES 22 challenge(https://isles22.grand-challenge.org/) for ischemic stroke segmentation.

Scope of the Study

Dataset Utilization

This study will primarily focus on the ISLES 2022 challenge dataset, which comprises 400 multi-vendor MRI cases. (Hernandez Petzsche et al., 2022) says "All imaging data are released in the native space without prior registration. Prior to release, skull-stripping was performed to de-identify patients."

Ensemble Model Development

Study will explore various ensemble learning techniques like bagging, boosting, stacking, and blending, these techniques combine all multiple models to produce better predictive results than using individual models.

Performance Evaluation

Research will consider multiple performance evaluation 11 metrics such as Accuracy), F1 Score, Area Under the ROC Curve (AUC-ROC), Precision, Log Loss, Recall (sensitivity Area Under the Precision-Recall Curve (AUC-PR), as each metric offers unique insights into different aspects of model behaviour.

Research Methodology

Dataset Selection and Preparation

The ISLES 2022 challenge dataset serves as the primary source for this study on ensemble learning for stroke segmentation. This dataset comprises 400 multi-vendor 6 MRI cases, split into a training set of 250 cases and a test set of 150 cases. (Hernandez Petzsche *et al.*, 2022)

The dataset includes four primary types of image files:

14 ADC (Apparent Diffusion Coefficient)

DWI (Diffusion Weighted Imaging)

FLAIR (Fluid Attenuated Inversion Recovery)

Ground-truth segmentation masks

All image files are provided in the NIfTI format (. nii.gz)

4 Data Preprocessing

Data preprocessing is a must do step to use the ISLES 2022 dataset for ensemble learning. "2 Prior to release skull-stripping was performed to de-identify patients".(Hernandez Petzsche *et al.*, 2022). Resizing images to a 7 range of 0-1 and Normalization of image will be performed to standardize the input across different MRI scanners and protocols. Image augmentation techniques may also 9 be applied to artificially expand the dataset and improve model generalization, Using NiBabel/SimpleITK library NIfTI image will be loaded and to access its data.

Base Model Development

The ensemble learning approach for Ischemic brain stroke segmentation will incorporate multiple individual segmentation algorithms, each individual segmentation models will serve as the base learners in the ensemble approach. Developing effective base models is essential for creating a robust ensemble learning system. By carefully selecting algorithms, properly preprocessing the data, rigorously training and evaluating models, and implementing appropriate ensemble techniques, powerful predictive system can be developed that leverages the strengths of multiple models.

Model Architecture Selection

This Phase appropriate model architecture will be selected suitable for segmentation tasks. Potential models may include:

Convolutional Neural Networks (CNNs): Particularly U-Net and its advanced variants.
Random Forests (RF): As a part of big data ensemble methods.
Gradient Boosting Techniques: Such as LightGBM and CatBoost.

Potential models to be included in the ensemble are U-Net, which is specifically designed for medical image segmentation tasks, and ResNet, known for its residual learning capabilities. Other advanced architectures like ResUNet++, NNU-net and efficient U-Net variants can also be integrated to enhance segmentation performance.

Implementation of Ensemble Learning Techniques

Bagging: Train multiple models using bagging techniques to enhance performance. Each model is trained on different subsets of the dataset with a focus on reducing variance.
Boosting: Utilize boosting techniques to iteratively train models that focus on previously misclassified instances. Modify weights of training examples after each model iteration to improve classification accuracy.
Stacking: Implement stacking by combining predictions from multiple base models and feeding them into a meta-learner model (e.g., logistic regression) for final predictions.

Performance Evaluation Metrics

To evaluate the model performance metrics such as Dice coefficient, accuracy, precision, recall, and area under the ROC curve (AU-ROC) for assessing segmentation quality, and also implement k-fold cross-validation techniques.

Requirements Resources

Data Accessibility

To conduct this research on ensemble learning for stroke segmentation, researchers must gain access to the ISLES 2022 dataset(https://isles22.grand-challenge.org/dataset/)

Computational Resources

Building and Training models through ensemble learning requires substantial computational power, primarily through high-performance GPUs. Specs should include a high number of CUDA cores to support parallel processing (minimum of 1,000 CUDA cores can support effective parallel processing in ensemble learning for ISLES segmentation, aiming for higher core counts greatly enhances performance and efficiency) and a memory capacity of at least 32GB, ensuring efficient handling of the large imaging datasets involved in the study. High memory bandwidth will facilitate faster data access during model training and evaluation.

Software Tools and Libraries

Implementation of ensemble learning models will necessitate the use of various software libraries. Essential tools include:

Deep Learning Tools - PyTorch, MONAI ((Medical Open Network for AI).

segmentation tools - nnU-Net, ITK-Snap and 3D Slicer.

Image processing libraries – ANTs, Elastix, NiftyReg, openCV

Ensemble learning libraries – ML-Ensemble, Pycobra

APPENDIX B – ISLES'22 Dataset

https://zenodo.org/records/6761864

The ISLES 2022 dataset consists of 400 MRI cases, carefully curated to represent a broad spectrum of ischemic stroke presentations. The dataset is structured as follows:

Training Set: 250 cases
Test Set: 150 cases

Sequences were chosen because they are the most important for radiologists when diagnosing subacute stroke lesions on MRI. This challenge is an effort to create and find top-performing methods. The dataset contains MRI scans collected from three different medical centres.

APPENDIX C - ETHICAL DECLARATION

I hereby declare that the research work presented in this thesis was carried out in accordance with the ethical standards and guidelines of Liverpool John Moores University. The study was conducted with integrity, transparency, and respect for academic and professional norms.

- **Originality:** This thesis is my own work and has not been submitted, in whole or in part, for any other degree or qualification. All sources of information and contributions from others have been duly acknowledged.
- **Data Usage:** The medical imaging data used in this research was obtained from publicly available datasets or authorized institutional sources. No personally identifiable patient information was accessed or disclosed.
- **Ethical Compliance:** Where applicable, the research adhered to ethical principles of non-maleficence, beneficence, and respect for privacy. The study did not involve direct human or animal experimentation.
- **Academic Integrity:** Plagiarism, fabrication, or falsification of data has not been practiced. All analyses, results, and interpretations are presented honestly and transparently.
- **Supervision and Approval:** The research was conducted under the supervision of faculty members and approved in accordance with the university's ethical review procedures.

I affirm that this thesis complies with the ethical requirements of postgraduate research and reflects my commitment.

PUNITH_N_FINAl_REPORT_THEISIS_Ensemble_learning_v8.6.d...

ORIGINALITY REPORT

65%	13%	10%	63%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Liverpool John Moores University Student Paper	62%
2	scholar.archive.org Internet Source	<1%
3	"MRI acute/sub-acute ischemic stroke segmentation with deep learning: A comprehensive review", Elsevier BV, 2026 Publication	<1%
4	Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dhirendra Kumar Shukla. "Recent Trends in Intelligent Computing and Communication", CRC Press, 2025 Publication	<1%
5	esource.dbs.ie Internet Source	<1%
6	lirias.kuleuven.be Internet Source	<1%
7	iaeme.com Internet Source	<1%
8	Sakshi Gupta, Aryan Chaudhary. "Artificial Intelligence in Healthcare - Trends, Applications, and Future Directions", Apple Academic Press, 2026 Publication	<1%
9	www.coursehero.com Internet Source	<1%
10	repository.effatuniversity.edu.sa Internet Source	<1%

11	revista.crcsc.org.br Internet Source	<1 %
12	Submitted to Universiti Teknologi MARA Student Paper	<1 %
13	Ezequiel de la Rosa, Mauricio Reyes, Sook-Lei Liew, Alexandre Hutton et al. "DeepISLES: a clinically validated ischemic stroke segmentation model from the ISLES'22 challenge", Nature Communications, 2025 Publication	<1 %
14	public-pages-files-2025.frontiersin.org Internet Source	<1 %
15	"Medical Information Computing", Springer Science and Business Media LLC, 2025 Publication	<1 %
16	Submitted to Monash University Student Paper	<1 %
17	S. N. Kumar, Sherin Zafar, Eduard Babulak, M. Afshar Alam, Farheen Siddiqui. "Artificial Intelligence in Telemedicine - Processing of Biosignals and Medical Images", CRC Press, 2023 Publication	<1 %
18	Tanmoy Das, Tanima Bhattacharya. "Evolution of Low-Dimensional Biosensors in Food Safety, Agriculture, and Healthcare", CRC Press, 2026 Publication	<1 %
19	arxiv.org Internet Source	<1 %
20	d-nb.info Internet Source	<1 %
21	fastercapital.com Internet Source	<1 %