Enhancing the Reliability of Medical Image Classification and Segmentation via Self-Supervised and Uncertainty-Aware Learning

by Guanqi Cheng

Master by Research

Supervised by: Principal Supervisor: Dr Ali Braytee
Co-Supervisor: Dr Mukesh Prasad

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney

March 2026

Certificate of Original Authorship

I, Guanqi Cheng, declare that this thesis is submitted in fulfilment of the requirements for the award of Master of Analytics (Research), in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution. This research was supported by an Australian Government Research Training Program (RTP) Scholarship doi.org/10.82133/C42F-K220.

Signature: Production Note: Signature removed prior to publication.

Date: 1/8/2025

Acknowledgement

Throughout this academic journey, I extend my most sincere gratitude to my principal supervisor, Dr. Ali Braytee, and co-supervisor, Dr. Mukesh Prasad. Their profound scholarly expertise, incisive academic guidance, and unwavering professional support have been pivotal in charting the intellectual trajectory and elevating the scholarly rigor of this research. Their meticulous feedback, adherence to scientific precision, and commitment to academic excellence have not only fostered my growth as an independent researcher but also instilled in me a deep appreciation for methodological rigor and intellectual integrity.

I am equally indebted to my esteemed colleagues in the research group for their collaborative spirit and collegial support. In particular, I would like to express special thanks to Mr. Xing Zi for his invaluable technical assistance and constructive discussions, which significantly contributed to the conceptual refinement and empirical implementation of this study. Our interdisciplinary dialogues during critical stages were instrumental in overcoming theoretical and methodological challenges, thereby enhancing the robustness of the research outcomes.

I would also like to express my sincere gratitude to the Institute of Brain Sciences at Shanghai University of Health for providing the datasets that formed the critical foundation for this research.

Finally, I must acknowledge the indispensable role of my family. Their unwavering emotional support, profound understanding of the demands of academic life, and steadfast belief in my scholarly endeavors have been the cornerstone of my perseverance. Their sacrifices and encouragement have enabled me to navigate the complexities of research while maintaining a balanced commitment to both academic excellence and personal growth.

Contents

Certificate of Original Authorship **ii**

Contents **iv**

List of Figures **viii**

List of Tables **viii**

List of Publications **ix**

Abstract **x**

1 Introduction **1**
1.1 Research Challenges 4
1.2 Research Objectives 7
1.3 Contributions to Knowledge 7
1.4 Research Significance 9
1.5 Thesis Organization 9

2 Literature Review **12**
2.1 The Class Imbalance Problem in Medical Imaging 13
2.1.1 Introduction to Class Imbalance 13
2.1.2 Overview of Self-supervised Learning (SSL) 13
2.1.3 Impact of Class Imbalance in Supervised Learning 14
2.1.4 Class Imbalance Approaches with Supervised Learning 15
2.1.5 Class Imbalance Approaches with Self-supervised Learning 16
2.2 Medical Image Segmentation in the Presence of Noisy Labels 18
2.2.1 Introduction to noisy labels in medical image segmentation 18
2.2.2 Approaches to Handle Label Noise in Segmentation 19
2.2.3 Instance-Dependent Transition Matrix (IDTM) Estimation and Applications of Uncertainty Estimation in Medical Imaging 21
2.3 Summary of Research Gaps 22

3 SCFD: A Self-Calibrated Feature Denoising Framework for Robust Medical Image Classification Under Extreme Class Imbalance **24**

3.1 Methodology . . . 25
3.1.1 A Brief on DINOv2 . . . 25
3.1.2 Self-Calibrated Feature Denoising (SCFD) Method . . . 26
3.1.2.1 Pre-trained DINOv2 Feature Extractor . . . 28
3.1.2.2 Self-Calibrated Auto-encoding for Minority Representation 28
3.1.2.3 Denoised Inference for Minority Classes . . . 30
3.2 Experiment . . . 32
3.2.1 Datasets . . . 32
3.2.2 Imbalance Settings . . . 33
3.2.3 Compared Metrics . . . 33
3.2.4 Implementation Details . . . 34
3.2.5 Computational Complexity Analysis . . . 34
3.3 Results . . . 35
3.3.1 Results on Lab-shanghai Dataset . . . 35
3.3.2 Results on Brain Tumor 4C Dataset . . . 37
3.3.3 Results on Chest CT Dataset . . . 38
3.3.4 Qualitative Visualization Results . . . 39

4 Label Uncertainty Transformation for Combating Noisy Labels in Medical Image Segmentation 42
4.1 Methodology . . . 43
4.1.1 Preliminaries . . . 44
4.1.2 IDTM Formulation and Noisy Label Modeling . . . 45
4.1.3 Uncertainty Estimation Ensemble Module (UEEM) . . . 45
4.1.4 Label Transformation Network (LTN) . . . 47
4.1.4.1 Distilled Sample Selection via Pixel Entropy . . . 47
4.1.4.2 IDTM Parameterization and Training Objective . . . 48
4.1.5 Target Segmentation Network (TSN) . . . 49
4.2 Experiment . . . 50
4.2.1 Datasets and Experimental Settings . . . 50
4.2.1.1 Datasets . . . 50
4.2.1.2 Noise Settings . . . 50
4.2.1.3 Compared Methods . . . 51
4.2.1.4 Evaluation Metrics . . . 51
4.2.2 Implementation Details . . . 52
4.3 Results . . . 52
4.3.1 Comparison of Lab Infarct Dataset . . . 52
4.3.2 Comparison with Brain Tumor Dataset . . . 54
4.3.3 Comparison of Merging Brain Tumor Dataset . . . 54
4.3.4 Ablation Experiment . . . 55
4.3.4.1 Impact of original LTN without UA Re-Weighting and pixel entropy regularizer . . . 56
4.3.4.2 The impact of pixel entropy regularizer(PE regularizer) . . 56
4.3.4.3 The impact of Uncertainty Re-weighting (URW) . . . 56

4.3.5 Pixel Entropy Regularizer Parameter Tuning 57

5 Conclusion and Future Work **64**

Bibliography **66**

List of Figures

2.1 Illustrative examples of simulated annotation noise. Each row shows an original ground-truth mask (Left) and its corresponding noisy mask (Right), demonstrating realistic label perturbations that mimic common human annotation errors. 19

3.1 The DINOv2 framework 26
3.2 Our proposed Self-Calibrated Feature Denoising (SCFD) framework 27
3.3 3D Visualization of Feature Embeddings via LDA with 5% Lesion Samples and Unchanged Normal Samples. 41

4.1 Our proposed Label Uncertainty Transformation (LUT) framework 46
4.2 Prediction results of our method's components in the ablation experiments on the Lab Infarct and Brain Tumor dataset. (a) Training samples. (b) Noisy labels. (c) Initial results by UEEM. (d) Predictive uncertainty. (e) Original LTN. (f) +Pixel entropy regularizer. (g) +URW 53
4.3 Dice and IoU performance across different values of the regularization parameter γ on the three datasets under 15% label noise. 58
4.4 Dice and IoU performance across different values of the regularization parameter γ on the three datasets under 30% label noise. 59
4.5 Dice and IoU performance across different values of the regularization parameter γ on the three datasets under 45% label noise. 60
4.6 Dice and IoU performance across different values of the regularization parameter γ on the three datasets under 60% label noise. 61
4.7 Dice and IoU performance across different values of the regularization parameter γ on the three datasets under 75% label noise. 62
4.8 Dice and IoU performance across different values of the regularization parameter γ on the three datasets under 90% label noise. 63

List of Tables

3.1 Summary of datasets used in our experiments. 32
3.2 Classification performance for Lab-shanghai dataset of various models evaluated at different numbers of lesions (100%, 50%, 20%, and 5%). The table reports Macro-avg Precision, Recall, and F1-score to assess model effectiveness across data scales. 35
3.3 Classification performance for the Brain Tumor 4C dataset of various models evaluated at different numbers of lesions (100%, 50%, 20%, and 5%). The table reports Macro-avg Precision, Recall, and F1-score to assess model effectiveness across data scales. 36
3.4 Classification performance for Chest CT dataset of various models evaluated at different numbers of lesions (100%, 50%, 20%, and 5%). The table reports Macro-avg Precision, Recall, and F1-score to assess model effectiveness across data scales. 37

4.1 Description of symbols 43
4.2 Details of the datasets used in the experiment. 50
4.3 Performance comparison of Lab Infarct dataset with different noise settings. The values are Dice and IoU scores. 52
4.4 Performance comparison of Brain Tumor dataset with different noise settings. The values are Dice and IoU scores. 53
4.5 Performance comparison of Merging Brain Tumor dataset with different noise settings. The values are Dice and IoU scores. 54
4.6 Ablation experiment on the Lab Infarct dataset with 15%, 45%, and 75% noisy data. 55
4.7 Ablation experiment on the Brain Tumor dataset with 15%, 45%, and 75% noisy data. 55
4.8 Ablation experiment on the Merging Brain Tumor dataset with 15%, 45%, and 75% noisy data. 55

List of Publications

Papers related to thesis

1. Cheng, G., Prasad, M., & Braytee, A. (n.d.). SCFD: A Self-Calibrated Feature Denoising Framework for Robust Medical Image Segmentation. Manuscript submitted to *The Proceedings of the International Joint Conference on Neural Networks 2026*. (Chapter 3)

2. Cheng, G., Zi, X., Prasad, M., & Braytee, A. (n.d.). Label uncertainty transformation for combating noisy labels in medical image segmentation. Manuscript under review at *ACM Transactions on Intelligent Systems and Technology*. (Chapter 4)

Abstract

Medical imaging is very critical in the diagnosis of different diseases such as cerebral infarction, brain tumors, and lung tumors. Though Artificial Intelligence (AI) has made great strides toward medical image analysis, several research challenges remain open. The accurate and broadly applicable AI models for medical image analysis are the determinants of success in clinical practice adoption. However, two major challenges still hinder progress. First, class imbalance is common in classification tasks, where pathological conditions occur much less frequently than normal cases. This disproportion often induces a bias in learning algorithms toward the majority class, consequently impairing the detection sensitivity for infrequent but clinically critical abnormalities. Such limitations may result in a delayed diagnosis or missed detection, which can adversely affect patient outcomes. Furthermore, the wide variability in the appearance of pathological findings complicates feature representation for minority classes, intensifying the imbalance issue. Second, label noise poses a major challenge in medical image segmentation datasets, especially in large-scale datasets or those annotated with limited or imprecise supervision. Annotation variability among experts, unclear anatomical boundaries, and the inherent difficulty of delineating specific lesions contribute to inconsistent and inaccurate labels. Human errors, including mislabeling or inconsistent boundary definitions, are frequent, especially in complex or subtle cases. This issue pertains to the deterioration of model performance caused by label noise, which undermines the effectiveness of training and results in suboptimal lesion localization and inaccurate boundary delineation in medical image segmentation. In clinical practice, such inaccuracies can compromise surgical planning, radiation therapy, and disease monitoring, where precise and reliable segmentation is crucial for effective treatment and prognosis.

This thesis tackles these challenges through two research approaches. The first method explores self-supervised learning, specifically utilizing DINOv2, an advanced vision transformer pre-trained model, to learn robust feature representations without reliance on labeled data. While not explicitly designed to address class imbalance, this approach can potentially mitigate its effects by improving the quality of learned representations in classification tasks. Results show that DINOv2 performs on par with supervised models while offering stronger representations for underrepresented classes. To further enhance feature clarity, a novel Self-Calibrated Feature Denoising (SCFD) framework is proposed. SCFD uses a lightweight autoencoder trained on minority-class features, regularized by the latent distribution of majority-class embeddings, to denoise and stabilize DINOv2

representations of minority samples, without modifying the DINOv2 backbone. Overall, this method consistently enhances classification performance across different levels of class imbalance.

The second method addresses label noise in medical image segmentation. Unlike traditional approaches that assume noise is random or class-independent, this study introduces a noise-robust framework termed Label Uncertainty Transformation (LUT). The proposed framework consists of three components. First, the Uncertainty Estimation Ensemble Module (UEEM) aggregates predictions from multiple models to quantify pixel-wise uncertainty. Based on these uncertainty estimates, a pixel-entropy-based sample selection mechanism identifies reliable distilled samples for subsequent training, effectively filtering out highly uncertain regions. Second, the Label Transformation Network (LTN) estimates an instance-dependent transition matrix. The network is optimized using an uncertainty-aware loss function combined with pixel entropy regularization, which jointly enhances the stability and accuracy of noise estimation, particularly improving modeling of sparse pathological structures such as lesions or tumors. Finally, the Target Segmentation Network (TSN) leverages the refined label information from UEEM and LTN to produce accurate segmentation predictions, resulting in improved robustness against noisy annotations. Our proposed methods were evaluated on distinct datasets corresponding to their tasks. For the classification task, experiments used three datasets, including a large-scale private cerebral infarction dataset (Lab-Shanghai) with 11,574 infarction and 11,574 normal cases, alongside two public datasets, Brain Tumor 4C and Chest CT from Kaggle. Performance metrics included the macro-average of Precision, Recall, and F1-score to address class imbalance. For the segmentation task, three datasets were employed: a private DWI-based infarction dataset (Lab Infarct) with data from 1,528 patients, and two public brain tumor datasets (Brain Tumor and Merging Brain Tumor) containing 3,064 and 2,642 images, respectively. Segmentation was evaluated using the Dice coefficient and Intersection over Union (IoU), demonstrating robustness and accuracy across different noise levels.

In summary, this thesis presents two rigorously developed and effective methodologies to address persistent challenges in medical image analysis. The first advances classification under class imbalance through self-supervised learning and feature denoising, while the second strengthens noise label segmentation robustness by modeling uncertainty and estimating an instance-dependent transition matrix. Together, these contributions aim to make automated medical imaging tools more reliable.

Chapter 1

Introduction

Medical risks such as brain tumors, cerebral infarction, and lung tumors pose serious threats to patient health and quality of life. Cerebral infarction is a major global health issue, marked by high incidence, fatality, and disability rates, often associated with risk factors such as hypertension, smoking, and diabetes [1, 2, 3]. Brain tumors, resulting from abnormal cell proliferation, can compress surrounding structures and impair brain function [4]. Malignant brain and spinal cord tumors constitute the majority of central nervous system (CNS) cancers and often progress rapidly with a poor prognosis [4, 5]. Meanwhile, lung cancer remains one of the most common and deadly cancers worldwide, contributing significantly to cancer-related mortality and is expected to impose a growing burden [6].

Medical images serve as detailed visualizations of the internal anatomical structures and physiological functions within the human body. They are acquired using various imaging techniques, including Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Ultrasound, and X-ray [7]. These imaging modalities are integral to clinical workflows, facilitating accurate diagnosis, informed treatment planning, and effective monitoring of disease progression.

Medical image classification refers to the task of assigning medical images to predefined categories based on their visual characteristics [8]. For example, in brain MRI scans,

classification algorithms may categorize images as either depicting healthy tissue or showing signs of pathological conditions such as tumors, strokes, or multiple sclerosis lesions. This automated categorization supports clinicians in diagnosis and treatment planning by highlighting critical abnormalities within the images [9]. This process is fundamental to diagnostic workflows, as it enables the identification of normal and pathological conditions from imaging data [10]. Accurate classification supports timely diagnosis, informs treatment decisions, and enhances patient management by providing consistent and objective evaluations. Accurate classification supports timely diagnosis, informs treatment decisions, and enhances patient management by providing consistent and objective evaluations. Given the growing volume and complexity of medical imaging data, automated classification methods powered by Artificial Intelligence (AI) and Machine Learning (ML) have become increasingly essential [11]. These techniques enable the development of predictive models that can efficiently analyze medical images, reduce diagnostic errors, and improve clinical workflow and outcomes.

Medical image segmentation involves identifying and outlining clinically relevant regions, such as organs, tissues, or lesions, in medical imaging data [12]. This process yields precise spatial localization that is critical for accurate diagnosis, treatment planning, and other clinical applications [13]. For instance, segmentation of brain magnetic resonance imaging (MRI) enables the delineation of tumor margins for surgical or radiotherapeutic planning and the identification of tissue types for monitoring disease progression [14]. In contrast to image-level classification, which assigns a single label to an entire image, segmentation produces pixel- or voxel-wise predictions that capture the precise morphology and spatial distribution of anatomical or pathological features [15]. The resulting fine-grained representations support more accurate diagnosis, facilitate individualized treatment strategies, and enable longitudinal assessment of disease, thereby contributing significantly to improved clinical outcomes [16].

In medical image classification, one major challenge is class imbalance, where the number of normal samples far exceeds that of certain disease cases, or vice versa [17]. This imbalance leads to biased models that favor majority classes while under-representing minority categories, thereby reducing diagnostic sensitivity [18].In medical classification tasks, extreme class imbalance occurs when the minority class constitutes only a minute proportion

of the dataset. A typical example can be found in brain tumor classification, where certain malignant subtypes—such as glioblastoma with specific molecular markers—appear far less frequently than benign lesions or normal tissue. This imbalance causes the model to favor predictions toward the dominant classes, impairing its ability to correctly identify rare but clinically significant cases, and ultimately reducing diagnostic sensitivity and reliability in high-stakes scenarios. Class imbalance is typically addressed using three main types of strategies: data-level techniques that manipulate the training distribution, algorithm-level modifications that adjust the learning process, and ensemble-based methods that combine multiple models to enhance robustness [19]. Although various methods have demonstrated notable progress in mitigating class imbalance, it remains an open and actively investigated challenge, particularly in the context of highly imbalanced and structurally complex datasets such as those encountered in medical imaging. Each approach category—data-level, algorithmic, or ensemble-based—has inherent limitations when applied to such demanding scenarios. Data-level methods aim to mitigate class imbalance by directly modifying the training data distribution, typically by oversampling minority classes or undersampling majority classes. Despite their simplicity and prevalence, these techniques may lead to overfitting—particularly in the case of duplicated or synthetically generated samples that lack anatomical plausibility—or to information degradation when relevant majority class instances are discarded [20]. Algorithm-level strategies, including class-weighted losses and margin-based adjustments, attempt to rebalance the optimization objective in favor of minority classes. Nonetheless, these methods are often sensitive to hyperparameter selection and do not explicitly address imbalances in the feature representation space, where minority class embeddings tend to be sparse, poorly clustered, or less discriminative [21]. Ensemble-based frameworks, which aggregate predictions across multiple learners, can mitigate class-specific biases but are typically associated with increased computational overhead and reduced interpretability [22].

Recent research has highlighted the potential of pretrained models and self-supervised learning (SSL) in addressing class imbalance [23]. Pretrained models are neural networks initially trained on large-scale datasets—labeled or unlabeled—to learn transferable feature representations, and then fine-tuned on downstream tasks to improve convergence and robustness [24]. SSL is a common pretraining strategy that leverages proxy tasks such as

context reconstruction or contrastive learning to learn class-agnostic features from unlabeled data [25]. These representations reduce intra-class variability and improve inter-class separability, particularly benefiting minority classes [26].

In contrast, medical image segmentation—crucial for localizing disease regions in conditions like brain tumors and cerebral infarction—faces the prominent challenge of label noise [27]. Label noise refers to inaccuracies or inconsistencies in training annotations, which are often caused by the difficulty and subjectivity of expert labeling [28, 29]. Moreover, factors such as small lesion size, ambiguous boundaries, and low contrast with surrounding tissue further complicate precise segmentation [30]. Training predictive models on noisy annotations can lead to overfitting and performance degradation [31], and despite extensive efforts, robust learning under label noise remains an open research problem in medical image segmentation [32].

1.1 Research Challenges

Despite significant advances in deep learning, AI models for medical image analysis continue to face critical challenges that limit their robustness and generalizability. Specifically, classification and segmentation tasks are affected by related but distinct issues: class imbalance and label noise, respectively. These problems substantially impede the development of reliable and clinically viable Artificial Intelligence (AI) models. The key research challenges are summarized as follows:

Class imbalance in medical image classification can lead to biased models that underperform on rare disease cases, reducing diagnostic sensitivity and producing unreliable predictions. This imbalance presents several critical challenges. Models often overfit to majority class patterns, leading to suboptimal feature representations and limited generalization for minority classes, which frequently correspond to rare but clinically significant conditions [33]. The scarcity of minority samples constrains the ability to capture diverse pathological variations, thereby undermining the robustness and reliability of classification systems in real-world clinical settings [34, 35]. Addressing the challenges posed by

such extreme imbalance is essential to developing effective and generalizable classification methods [36].

Moreover, extreme imbalance complicates model calibration, causing overconfident predictions on the majority classes and unreliable, ambiguous outputs on minority classes, which undermines clinical trust and decision-making. Class imbalance often occurs alongside other challenges, such as label noise (incorrect or uncertain annotations) and intra-class heterogeneity (large variation within the same class). These combined factors make it even harder for models to learn accurate and reliable classifications [37, 38]. These intertwined issues make it challenging to design algorithms that are both robust to imbalance and stable during training, leaving this as a prominent open problem in the field.

Current approaches to mitigate class imbalance include data-level methods such as oversampling and undersampling to balance class distributions, though these may introduce overfitting or information loss. Algorithm-level techniques employ cost-sensitive learning and adaptive loss functions like focal loss to emphasize minority classes during training. Recent advances incorporate dynamic reweighting schemes guided by sample difficulty or uncertainty to improve robustness, as well as self-supervised and contrastive learning paradigms to enhance feature representation under imbalance [36, 37, 38, 39]. Despite recent advances, achieving a principled balance between representation robustness, prediction calibration, and training stability under severe class imbalance remains a significant open problem. In particular, the under-representation of minority classes often leads to dispersed and non-discriminative feature embeddings, miscalibrated confidence estimates, and unstable convergence behavior. Although self-supervised learning and pretrained models can provide more structured and transferable representations—especially in data-scarce scenarios—their integration into supervised pipelines is non-trivial. Pre-trained features may not align with task-specific decision boundaries under imbalance, and their adaptation can introduce gradient mismatch, feature drift, or overfitting when fine-tuning on limited minority data. These challenges highlight the need for more targeted methods to reconcile pretrained representations with imbalance-aware optimization objectives. These challenges motivate the investigation of learning strategies that reduce reliance on large-scale, clean, and well-balanced annotations. In medical imaging, acquiring accurate labels, particularly for rare disease categories, is costly and often affected by inter-observer

variability, which further amplifies class imbalance and label noise. Self-supervised learning provides an effective alternative by leveraging unlabeled data to learn transferable and semantically meaningful representations that can be used to initialize downstream classification models. By decoupling representation learning from noisy and imbalanced supervision, self-supervised approaches can enhance feature discriminability, robustness, and calibration in rare disease classification. This leads to the following research question.

- **RQ1: How can self-supervised learning help address class imbalance in medical classification tasks?**

Label Noise in Medical Image Segmentation. Label noise in medical image segmentation presents significant challenges beyond mere annotation errors [40]. Noisy labels often introduce misleading spatial information, causing models to overfit to incorrect or inconsistent regions, which severely degrades segmentation accuracy and generalization. For example, in tumor segmentation tasks, inaccurate manual annotations due to unclear tumor boundaries or inter-observer variability can lead the model to learn erroneous tumor shapes, resulting in poor delineation and unreliable clinical assessments. The high dimensionality and complex anatomical variability in medical images amplify these difficulties, making noise correction and mitigation particularly challenging [41]. Existing noise-robust techniques for medical image segmentation primarily rely on label correction, loss reweighting, and uncertainty modeling [42]. Label correction approaches iteratively refine noisy annotations based on model feedback or prior knowledge; however, these methods often struggle to accurately distinguish true lesion boundaries from annotation errors, especially in small or low-contrast regions [43]. Loss reweighting schemes aim to mitigate the influence of unreliable samples by adjusting their contribution during training, but may inadvertently down-weight challenging yet clinically relevant regions, potentially impairing model sensitivity [44]. Uncertainty modeling seeks to estimate prediction confidence to guide training, yet the spatial granularity of such uncertainty estimates is frequently insufficient to resolve fine boundary ambiguities arising from annotation inconsistencies [45]. These challenges underscore the need for more precise and robust strategies that can effectively separate genuine pathological features from label noise, thereby improving segmentation reliability in complex clinical contexts. Furthermore, the scarcity of clean annotated data and the

high cost of obtaining expert consensus hinder the development of reliable benchmarks and validation protocols. Consequently, designing segmentation algorithms that can inherently tolerate or adapt to noisy labels, while preserving anatomical fidelity and clinical relevance, remains an unresolved and critical research challenge [31, 32, 46]. Based on the above research issues, our research question is as follows:

- **RQ2: How can the impact of label noise be minimized in medical segmentation tasks?**

1.2 Research Objectives

The objectives of this research are to improve the reliability and performance of medical image analysis in two challenging scenarios:

- Enhancing classification performance on minority classes through self-supervised learning by increasing the model's sensitivity to underrepresented patterns.
- Improving training data quality by identifying and correcting noisy labels, allowing the model to learn from reliable examples and achieve more accurate segmentation.

1.3 Contributions to Knowledge

This subsection presents two contributions to knowledge based on the two research questions above.

- A lightweight denoising and regularization scheme is proposed, operating on intermediate feature representations to enhance minority class discrimination without modifying the backbone network. The method exhibits consistent performance across a wide range of architectures and imbalance levels.

- An uncertainty-guided framework is designed to mitigate label noise by combining ensemble-based predictive variance and pixel-wise entropy, which stabilizes the estimation of the transition matrix and improves segmentation accuracy under noisy supervision.

Contribution 1: Self-Calibrated Feature Denoising (SCFD) and Robust Evaluation Across Diverse Backbones A novel Self-Calibrated Feature Denoising (SCFD) framework is proposed to enhance minority class feature representations by applying denoising and regularization directly on intermediate feature maps without requiring backbone retraining. This reduces computational costs while enhancing class discrimination under both moderate and extreme class imbalance. The advantages of self-supervised learning (SSL) become particularly evident in highly imbalanced scenarios, making it a promising approach. The framework's scalability and robustness are demonstrated through extensive evaluations on 22 popular backbone architectures—including convolutional and transformer-based models—under varying class imbalance ratios. SCFD consistently achieves superior accuracy and minority class recall compared to baseline methods, showcasing broad applicability and resilience in real-world imbalanced scenarios. This contribution addresses RQ1 by effectively improving feature quality for underrepresented classes through an efficient plug-in module.

Contribution 2: Uncertainty-Guided Sample Distillation and Stable IDTM Estimation with Extensive Experimental Validation A deep ensemble-based uncertainty estimation network, combined with pixel-entropy category weighting, is introduced to identify and select refined distilled samples that reduce the impact of label noise. Additionally, an uncertainty re-weighting strategy and pixel entropy regularization are employed to stabilize the Instance-Dependent Transition Matrix (IDTM) estimation, improving robustness in noisy and imbalanced contexts. Comprehensive experiments on three benchmark datasets with multiple model variants validate the effectiveness of these approaches, demonstrating significant improvements in segmentation accuracy and label-noise mitigation compared with state-of-the-art methods. This contribution addresses RQ 2 by effectively mitigating the adverse effects of noisy labels through uncertainty-guided

sample selection and robust IDTM estimation, thereby improving model reliability and segmentation performance in challenging medical imaging scenarios.

1.4 Research Significance

This thesis proposes novel algorithms that address two critical challenges in medical image analysis: label noise in segmentation and class imbalance in classification. Specifically, uncertainty estimation combined with the estimation of the Instance-Dependent Transition Matrix (IDTM) effectively mitigates the impact of label noise, enhancing segmentation robustness and accuracy. Meanwhile, feature denoising techniques target class imbalance by improving the representation of minority classes in classification tasks. Extensive validation across multiple datasets and model architectures demonstrates the scalability, adaptability, and generalizability of the proposed framework, advancing computational methodologies in medical imaging.

Clinically, these methods enable more precise delineation of pathological regions despite annotation inconsistencies and improve sensitivity in detecting rare but clinically significant conditions. By strengthening feature representations without requiring extensive retraining, the framework supports applications across diverse clinical datasets and resource-limited environments. Together, these advances facilitate the development of reliable medical imaging tools ready for integration into clinical workflows. The effectiveness of the proposed approaches is demonstrated through experiments presented in Chapters 3 and 4.

1.5 Thesis Organization

Each chapter of the thesis is composed as follows:

- Chapter 1: This chapter outlines the key research challenges and underlying motivations that drive this study. It further defines the specific research questions addressed and summarizes the main contributions made, including a list of related publications.

- Chapter 2: This chapter reviews prior works related to medical image analysis, particularly focusing on techniques relevant to classification under class imbalance and segmentation under label noise. We discuss advances in self-supervised and supervised pre-trained models, class imbalance in medical datasets, learning from noisy segmentation labels, and recent efforts in estimating the instance-dependent transition matrix. We conclude with a summary of research gaps that motivate our study.

- Chapter 3: This chapter introduces a novel strategy for robust medical image classification under extreme class imbalance. Our approach builds upon DINOv2, a powerful self-supervised visual encoder. It incorporates a Self-Calibrated Feature Denoising (SCFD) framework to selectively refine noisy minority class representations without modifying the pretrained backbone. It consists of three key stages: (1) feature extraction using a fixed DINOv2 encoder, (2) minority-specific denoising via a lightweight autoencoder, and (3) denoised inference for enhanced classification robustness. A large number of experiments across three datasets have demonstrated the effectiveness of our method for class-imbalanced classification.

- Chapter 4: This chapter addresses the label noise problem of brain tumor and cerebral infarction segmentation by proposing a Label Uncertainty Transformation (LUT) framework that utilizes prediction uncertainty to identify pixels prone to label noise. The key objective is to parameterize the Instance-Dependent Transition Matrix (IDTM). Initially, an Uncertainty Estimation Ensemble Module generates preliminary segmentation results and uncertainty maps, from which high-confidence distilled samples are extracted. These samples are refined using pixel-entropy-based class weighting and then used to train a Label Transformation Network. This network estimates the IDTM to transform noisy predictions into clean labels. To facilitate effective training, pixel entropy regularization and uncertainty re-weighting strategies are introduced to mitigate the effect of mislabeled pixels. Upon completion of training, the Target Segmentation Network produces the final segmentation outputs. A large number of experiments across three datasets have demonstrated the effectiveness of our method for label noise segmentation.

- Chapter 5: This chapter presents the conclusion of the thesis, summarizing the main findings and contributions, as well as outlining future research directions.

Chapter 2

Literature Review

This chapter reviews the literature relevant to this thesis. Section 2.1 focuses on the class imbalance problem in medical imaging. It begins by defining class imbalance with practical examples and then reviews existing strategies for handling imbalanced data, including data-level, algorithm-level, and loss-level approaches. Additionally, this section discusses self-supervised learning (SSL) methods and their potential to mitigate class imbalance by leveraging unlabeled data. Section 2.2 surveys the literature on medical image segmentation with noisy labels. It first introduces the concept and causes of label noise in medical segmentation, then reviews approaches for handling noisy labels. Finally, it covers uncertainty estimation techniques used to identify and mitigate the impact of noisy annotations. Research gaps are highlighted within each section to motivate the methods proposed in this thesis. Research gaps are highlighted throughout these sections to motivate the methods proposed in this thesis. Section 2.3 summarizes the key limitations in the existing literature and provides a comprehensive discussion of the open challenges addressed by this work.

2.1 The Class Imbalance Problem in Medical Imaging

2.1.1 Introduction to Class Imbalance

Class imbalance frequently occurs in medical imaging, where abnormal or pathological regions represent only a small fraction of the overall data compared to normal tissues [35]. This imbalance tends to cause machine learning models to favor the majority class, which in turn leads to poor performance on detecting or segmenting the minority classes that are often clinically critical [37].

To mitigate this issue, researchers have generally adopted three types of strategies. The first involves adjusting the data distribution itself through methods such as oversampling minority classes, undersampling majority classes, or generating synthetic samples to balance the dataset [47]. The second strategy focuses on modifying the learning algorithm, often by assigning higher costs or weights to errors made on minority classes, encouraging the model to pay more attention to these classes [48]. The third category encompasses designing or adapting loss functions that emphasize difficult or rare samples, examples being focal loss or Dice loss [49]. These approaches can be used individually or in combination to enhance model robustness against class imbalance.

2.1.2 Overview of Self-supervised Learning (SSL)

Self-supervised learning (SSL) has emerged as an effective approach for leveraging large amounts of unlabeled data by designing proxy tasks that encourage models to learn meaningful representations without relying on manual annotations [50]. This paradigm is particularly valuable in medical imaging, where obtaining reliable labels is often costly and time-consuming. Common SSL techniques include contrastive learning, masked image modeling, and self-distillation, all aimed at capturing underlying semantic and structural information from the data itself [51, 52].

When applied to imbalanced datasets, SSL shows promise by enabling the model to learn more generalizable features that are less biased toward majority classes. Several methods

have incorporated balanced sampling strategies, clustering, or hybrid frameworks that combine SSL with supervised learning to improve minority class representation [53, 54].

However, challenges remain when deploying SSL on highly imbalanced medical datasets. Many existing approaches assume relatively balanced data distributions, and their performance can deteriorate when minority classes are severely underrepresented. Although recent state-of-the-art pretrained self-supervised learning (SSL) frameworks such as DINOv2 [55] demonstrate strong generalization across various visual tasks, their ability to explicitly mitigate class imbalance—especially within the medical imaging domain—has yet to be thoroughly examined. This limitation highlights the need for further investigation into SSL methods tailored to address class imbalance more effectively.

2.1.3 Impact of Class Imbalance in Supervised Learning

Supervised learning (SL) relies on labeled data to model the relationship between inputs and outputs [56]. As the learning process is guided entirely by the provided annotations, any imbalance in the label distribution directly affects how the model learns to distinguish between classes. In clinical image datasets, pathological findings such as malignant tumors, infarcts, or lung lesions are typically underrepresented compared to normal cases [57]. This imbalance often causes models to prioritize the majority class during training, while failing to adequately learn patterns associated with rare but clinically important categories [19].

Unlike self-supervised approaches, SL can only learn from explicitly labeled data [28]. When certain classes are poorly represented in the training data, the model receives limited exposure to their features, leading to biased decision boundaries. Although such models may achieve high overall accuracy by frequently predicting the dominant class, they often exhibit low sensitivity to minority classes [36]. In medical applications, this can result in missed or delayed diagnoses, particularly when detecting rare diseases [58].

The problem is further exacerbated in multi-class classification or when class differences are subtle, as in the detection of early-stage diseases or the identification of rare subtypes [59]. Here, the scarcity of labeled examples from minority classes restricts the model's ability to capture intra-class variability, leading to unreliable or inconsistent predictions [60].

As a result, the inherent label dependence of supervised learning makes it particularly vulnerable to class imbalance, posing a major challenge in developing reliable and clinically applicable classification systems.

2.1.4 Class Imbalance Approaches with Supervised Learning

Rajendran et al. [61] proposed a breast cancer prediction pipeline that combines data selection, preprocessing, class balancing methods, and several supervised classifiers, including Bayesian Networks, Random Forests, and Decision Trees. While the method led to improvements in overall classification accuracy, its ability to detect minority cases—specifically, cancer diagnoses—was limited under severe class imbalance. It also lacked specific mechanisms for identifying high-risk clinical cases. Liu et al. [35] developed an ensemble learning approach that applies SMOTE over-sampling and CVCF noise filtering, using multiple SVM classifiers with different kernel functions. Classifier weights were optimized using the SAGA algorithm. Although performance on imbalanced datasets improved, the method remained sensitive to noisy data and showed limited generalization when applied to highly skewed distributions. Similarly, Hu et al. [62] introduced a supervised deep learning model with a self-adaptive auxiliary loss (DSN-SAAL), which re-weights cross-entropy using the effective number of samples and adds reverse cross-entropy regularization. This design aimed to improve COVID-19 classification from imbalanced and noisy CT images. However, jointly addressing class imbalance and label noise remains a challenge, often leading to biased predictions and reduced performance on minority samples. Esposito et al. [63] proposed GHOST (Generalized tHreshOld ShifTing), an automatic threshold optimization method that selects class-specific thresholds to maximize Cohen's kappa on the training set, thereby improving classification performance under class imbalance. However, the method relies heavily on the predicted probability distributions from the training data, making it sensitive to label noise and potentially less effective in scenarios with severe class imbalance, where minority class detection remains challenging. Rehman et al. [64] proposed a stroke prediction framework that integrates data preprocessing, ADASYN and SMOTE oversampling, and ensemble learning (stacking and soft voting classifiers) to improve classification performance on imbalanced stroke datasets. While the approach enhances minority class detection, it relies heavily on oversampling techniques that may

introduce synthetic noise and increase the risk of overfitting, particularly when applied to small or noisy datasets, potentially limiting its generalizability in real-world clinical settings. Pandey et al. [65] proposed a two-level framework that systematically integrates three data-level techniques (RUS, ROS, and SMOTE) with two algorithm-level approaches (cost-sensitive learning and multiple machine learning classifiers) to effectively address class imbalance and improve the prediction of positive heart disease cases. While this method offers a comprehensive combination of data-level and algorithm-level strategies, it faces several limitations, including the risk of overfitting, potential information loss, difficulty in defining an appropriate cost matrix, and limited generalization ability. Qu et al. [66] employed Random Forest and transfer-learned ResNet-18 CNN models to investigate the impact of class imbalance and various data-level mitigation techniques—such as augmentation, undersampling, and oversampling—on the performance of multiclass and binary medical image classification tasks. While these methods effectively simulate and address class imbalance, they carry risks of overfitting due to data augmentation and may suffer from limited generalizability due to the narrow diversity of the datasets used and the absence of algorithm-level adjustments, such as class weighting, particularly in CNN models.

2.1.5 Class Imbalance Approaches with Self-supervised Learning

Elbatel et al. [67] proposed Fed-MAS, a federated learning framework that incorporates self-supervised priors to estimate class-aware divergence and adjust model aggregation weights. The method showed improvements in learning from minority classes in imbalanced medical imaging tasks. Still, many federated approaches struggle to capture class-level differences and often generalize poorly under strong imbalance and non-IID conditions. Su et al. [68] introduced STGAN, a two-stage GAN-based augmentation framework that uses self-supervised learning and Freeze-D to transfer both shared and class-specific features, thereby generating synthetic skin lesion images to support classification under imbalanced conditions. Despite improved performance, the limited availability of labeled data and significant class imbalance continued to affect accuracy on underrepresented categories. Li et al. [69] proposed a dual-track self-supervised learning-to-rank model that selects diverse and difficult negative CT samples for more effective training of COVID-19

classifiers under imbalance. This approach addressed some of the limitations of earlier methods, such as inefficient sample selection and high computational cost. Liu et al. [70] propose SSLDTI, a novel drug–target interaction prediction model that addresses class imbalance by integrating a positive samples compensation coefficient into the loss function and leveraging both contrastive learning and a self-prediction mechanism to capture true interaction representations effectively. This approach improves the model's sensitivity to minority class samples and enhances prediction accuracy for rare interactions. However, despite these advances, SSLDTI may still face challenges in learning complex interaction patterns when positive samples are extremely scarce, which can limit its generalization performance in highly imbalanced settings. Chen et al. [71] proposed a self-supervised learning framework for addressing class imbalance in medical image classification. The method first performs unsupervised pretraining using techniques such as Momentum Contrast (MoCo) or Rotation Prediction to extract generalizable representations from unlabeled data. This is followed by a supervised fine-tuning phase using Balanced-MixUp, which interpolates between minority- and majority class samples to improve learning for the minority class. While the approach enhances feature representation for underrepresented classes, it still relies on the quality of labels and sampling strategies in the supervised stage. Moreover, the self-supervised phase does not explicitly address class imbalance, limiting its effectiveness in cases with extremely imbalanced or long-tailed distributions. Azizi et al. [72] proposed a self-supervised learning framework that combines SimCLR pretraining with Multi-Instance Contrastive Learning (MICLe) to learn robust representations from large-scale unlabelled medical images. This approach enables effective transfer learning to downstream classification tasks, particularly under class-imbalanced conditions. While MICLe improves feature robustness and enhances minority class recognition, the method still relies on a sufficient number of instances per class and incurs significant computational cost during pretraining, which may limit its performance in extremely low-shot or highly imbalanced scenarios. Mirzaee et al. [73] present a multi-label chest X-ray classification framework that effectively tackles class imbalance and disease co-occurrence by using Adaptive Class Balance Augmentation to dynamically adjust data augmentation based on class frequencies and their co-occurrences. The approach combines attention-based contrastive self-supervised learning for robust feature extraction, an ensemble of binary

classifiers, a correlation learning module integrating statistical and biological disease relationships, and a focal weighted cross-entropy loss to emphasize hard-to-classify examples. While this comprehensive method improves accuracy and robustness, it involves high computational complexity and numerous hyperparameters, making training resource-intensive and tuning challenging. Additionally, dependence on predefined disease co-occurrence matrices may reduce its ability to generalize to rare or novel disease combinations.

2.2 Medical Image Segmentation in the Presence of Noisy Labels

2.2.1 Introduction to noisy labels in medical image segmentation

In medical image segmentation, label noise refers to inaccuracies or inconsistencies in the annotated ground truth, often arising when the assigned labels do not precisely reflect the underlying anatomical or pathological structures [29]. Noisy labels can manifest as incorrect boundaries, misclassified regions, or completely erroneous annotations, all of which negatively affect model training and generalization [32].

Such label noise is particularly prevalent in the medical domain for several reasons. First, manual annotation of medical images is highly subjective and depends heavily on expert interpretation, which can vary across clinicians due to differences in experience, fatigue, or ambiguity in the image itself [46]. Second, the annotation process is time-intensive and often constrained by limited expert availability, leading to reliance on weak supervision, semi-automatic tools, or outsourcing—each of which introduces potential inaccuracies [74]. Third, certain pathological features, such as small lesions or diffuse boundaries, are inherently difficult to delineate precisely, increasing the likelihood of annotation errors [75].

These challenges make label noise a significant barrier to developing reliable and robust deep learning models for medical image segmentation. Addressing this issue requires techniques that are not only robust to annotation errors but also capable of distinguishing between systematic noise and true signal in the data.

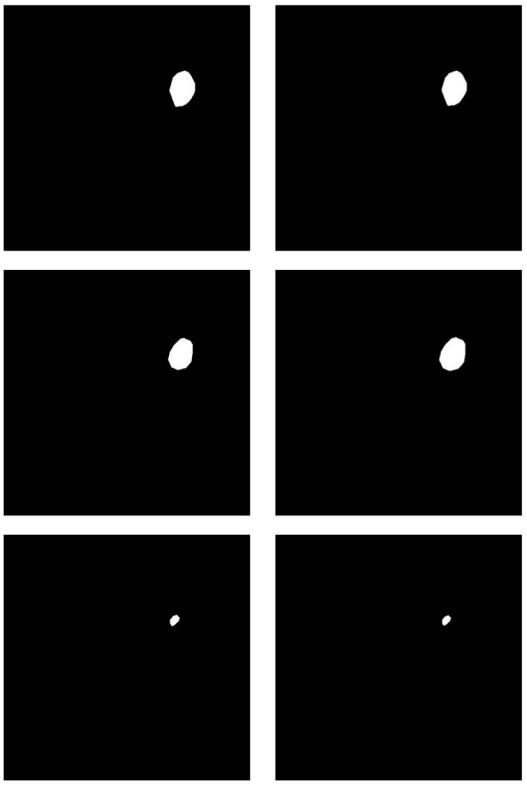


FIGURE 2.1: Illustrative examples of simulated annotation noise. Each row shows an original ground-truth mask (Left) and its corresponding noisy mask (Right), demonstrating realistic label perturbations that mimic common human annotation errors.

2.2.2 Approaches to Handle Label Noise in Segmentation

Segmentation masks serve as essential supervision signals in medical image segmentation, directly affecting model performance [76]. However, pixel-level annotation is costly and prone to errors, leading to noisy labels where, for example, healthy tissue may be mislabeled as tumor and vice versa [29]. To tackle this, Zhang et al. [77] introduced a teacher-student framework that integrates Confident Learning (CL) and SLSR to correct noisy annotations, although their approach relies on precise noise estimation.

Felfeliyan et al. [78] applied soft labels with a noise-robust loss in weakly supervised segmentation, yet soft labels often struggle to preserve accurate boundaries. Liu et al. [79] proposed an RSF-assisted label correction method incorporating image features, but its success hinges on the quality of the input images. Wu et al. [80] designed a two-stage

weakly supervised pipeline that uses size constraints and noise-resistant losses; nonetheless, the resulting pseudo-labels may still contain noise. Shi et al. [81] developed a resilient learning framework leveraging uncertainty estimation for 3D segmentation, but overly aggressive filtering risks discarding useful information.

Li et al. [82] combined uncertainty and reliability estimation with contrastive learning to refine noisy labels and improve brain segmentation. However, their method requires teacher-student discrepancy measurements and a meta-validation set, which increases computational burden and relies heavily on the quality of validation data, thus limiting scalability to larger or uncurated datasets.

Guo et al. [83] proposed the Collaborative Learning with Curriculum Selection (CLCS) framework, which employs a dual-branch network combined with curriculum dynamic thresholding and collaborative confidence voting to effectively distinguish clean and noisy labels. The method also introduces a noise balance loss to balance convergence on clean labels and robustness against noisy ones, thereby improving the learning performance in medical image segmentation with noisy annotations. However, the framework's complexity, reliance on multi-stage training schedules, and dynamic adjustments in early training may lead to instability and increased computational costs.

Wang et al. [84] proposed a method based on the Mean Teacher framework that integrates a shallow denoising module and a soft correction loss. By leveraging the teacher model to correct noisy labels and incorporating uncertainty estimation, their approach effectively improves segmentation performance on small lesions in medical images with noisy annotations. However, the method involves a complex architecture with multiple modules and hyperparameters, is sensitive to the teacher model's quality and the design of the soft correction strategy, and incurs high computational costs during training.

Penso et al. [85] proposed a noise-robust confidence calibration method that combines label noise matrix estimation with temperature scaling to effectively adjust the model's output probability distribution, achieving more accurate confidence estimates on noisy labeled data. However, the method relies heavily on accurate estimation of the label noise matrix and assumes a strong noise distribution, which may limit its performance in real-world scenarios with highly imbalanced classes or complex noise patterns.

2.2.3 Instance-Dependent Transition Matrix (IDTM) Estimation and Applications of Uncertainty Estimation in Medical Imaging

The Instance-Dependent Transition Matrix (IDTM) [86] models the probability of categorical mislabeling by incorporating instance-specific features, and has been widely used to address label noise in image classification. Cheng et al. [87] proposed training a robust classifier by jointly learning the IDTM alongside manifold regularization to mitigate the impact of label noise. Similarly, Zhu et al. [88] introduced a Covariance-Assisted Learning (CAL) method that leverages second-order statistics to tackle instance-dependent label noise (IDN). Li et al. [89] proposed a graph-based knowledge transfer framework to improve classification performance under noisy annotations. The approach extracts global noise patterns and constructs an annotator similarity graph to facilitate knowledge transfer via graph convolution, enabling the learning of annotator and instance-dependent transition matrices. Wang et al. [90] propose a method to learn a statistically consistent classifier from noisy training data by adaptively estimating an instance-dependent noise transition matrix (IDNT). The approach begins by separating the dataset into clean and noisy subsets using a noise-aware module that models sample losses with a two-component Gaussian Mixture Model, leveraging the memorization properties of deep neural networks. An instance-label confusion module then computes similarity scores between image features and label embeddings to form an instance-label confusion matrix (ILCM), capturing the relationships between instances and categories. Based on these components, the IDNT is estimated by combining the ILCM with either the provided label or the network's prediction, with the weights applied to each sample based on its estimated noise probability. Finally, the classifier is trained using an F-Correction loss that integrates the estimated noise transition matrix to mitigate the impact of label noise.

Uncertainty estimation has recently become a key focus in medical imaging research, particularly in segmentation and classification tasks, where it aids in model calibration, sample selection, and quality control [91]. Mehrtash et al. [92] computed segment-level predictive uncertainty by averaging pixel-wise entropy over predicted foreground regions, yielding a single confidence score per segment. While effective, this approach may underestimate uncertainty by ignoring spatial correlations and structural dependencies within irregular

or fragmented segments. Judge et al. [93] proposed learning a joint latent space for images and segmentation maps to estimate uncertainty based on similarity in the latent representation. However, this method relies heavily on the quality of the latent space and incurs computational overhead from multiple comparisons of decoded samples. Ju et al. [94] developed a dual-uncertainty estimation framework that combines an improved Direct Uncertainty Prediction (iDUP) to handle multi-annotator disagreement with Monte Carlo Dropout-based predictive uncertainty weighting for single-label annotations. By integrating sample selection and curriculum training, their approach enhances robustness under noisy label conditions. Nevertheless, it is sensitive to annotation quantity and quality, entails high computational cost, and relies on threshold-based filtering and hyperparameter tuning, which may result in the loss of minority class samples and increased training complexity. Sherkatghanad et al. [95] proposed a method that improves prediction accuracy and robustness by leveraging Bayesian Model Averaging (BMA) over multiple predictions generated via test-time augmentation (TTA). Their approach adaptively selects models based on likelihood measures, integrates diverse augmented inputs to produce a consensus output, and estimates predictive uncertainty. This framework effectively combines ensemble learning with probabilistic model averaging to enhance reliability during testing. Zhou et al. [96] propose a novel encoder-decoder framework that integrates hierarchical feature fusion, uncertainty-guided cross-level fusion, and scale aggregation modules to enhance lesion segmentation in medical images. By effectively combining multi-scale and multi-level features and leveraging uncertainty to guide feature integration, their method achieves improved accuracy and robustness. However, the model's complexity and the additional computational overhead introduced by uncertainty estimation may limit its efficiency in real-time clinical applications.

2.3 Summary of Research Gaps

Considerable progress has been made in addressing class imbalance in medical image classification. Existing techniques, including data resampling, cost-sensitive learning, and specialized loss functions, have demonstrated effectiveness under controlled or moderately imbalanced conditions. However, their performance often degrades in real-world clinical

settings, where class distributions are often highly skewed and heterogeneous. Although self-supervised learning (SSL) methods offer promising capabilities by leveraging unlabeled data to learn generalizable representations, their capacity to explicitly address severe class imbalance remains limited. In particular, current SSL pre-trained models lack mechanisms specifically designed to mitigate the adverse effects of imbalance.

Label noise presents a parallel challenge in medical image segmentation. Methods such as teacher–student frameworks, uncertainty-based filtering, and soft-label regularization have shown potential for alleviating noisy annotations. Nonetheless, these approaches frequently depend on assumptions—such as the accuracy of pseudo-labels or the reliability of uncertainty estimation—that may not hold in complex, weakly supervised scenarios. Instance-Dependent Transition Matrices (IDTMs) offer a principled framework for modeling structured label noise, yet their application to dense prediction tasks like pixel-wise segmentation remains underexplored. Moreover, many noise-robust approaches demand significant computational resources or rely on strong supervision, thereby limiting their scalability to large-scale, sparsely annotated datasets.

Importantly, class imbalance in classification and label noise in segmentation have conventionally been studied as independent issues. However, their frequent coexistence in clinical practice highlights the need for unified frameworks that can address both challenges concurrently, thereby improving model robustness and reliability in real-world medical imaging workflows.

Chapter 3

SCFD: A Self-Calibrated Feature Denoising Framework for Robust Medical Image Classification Under Extreme Class Imbalance

This chapter introduces a set of approaches to address the class imbalance problem in binary and multi-class classification tasks on medical image datasets. Class imbalance occurs when the number of instances in one class significantly exceeds that in another, a common issue in medical applications due to the rarity of certain conditions or the high cost and complexity of data collection. This imbalance often causes conventional classification models to be biased toward the majority class, resulting in poor detection performance for the minority class, which is often of greater clinical interest.

Despite numerous proposed solutions in the literature, no universally optimal method exists for handling class imbalance across all contexts. In particular, representation learning models often fail to capture meaningful patterns for minority classes, especially in high-dimensional or low-sample scenarios. This chapter addresses these challenges by proposing a novel feature-level denoising strategy that can be integrated into state-of-the-art representation-learning backbones.

Specifically, three key contributions are introduced in this chapter. First, we incorporate a strong self-supervised representation model (DINOv2) into the binary classification pipeline for medical imaging. Second, we propose a Self-Calibrated Feature Denoising (SCFD) module designed to enhance minority class representations through denoising and feature regularization. Third, we present comprehensive experimental results across a variety of imbalance ratios and backbone networks, demonstrating the robustness and effectiveness of our proposed framework.

The rest of this chapter is organized as follows: Section 3.1 introduces the SCFD-enhanced classification framework based on DINOv2. Subsequent sections describe the experimental setup, evaluation metrics, and performance comparisons with state-of-the-art methods under varying degrees of class imbalance.

3.1 Methodology

In this section, we introduce our novel framework, Self-Calibrated Feature Denoising (SCFD), designed to enhance minority class representations in class-imbalanced medical image classification. SCFD employs a minority-specific autoencoder regularized by a majority class before effectively denosing features, thereby improving robustness under extreme class imbalance. Before detailing SCFD, we provide a brief overview of DINOv2, the powerful self-supervised vision transformer backbone that underpins our framework.

3.1.1 A Brief on DINOv2

DINOv2 is a self-supervised model developed by Meta that learns high-level semantic representations from large-scale unlabeled data without manual annotations [97, 98, 99]. Fig. 3.1 shows the DINOv2 architecture. It achieves strong transferability across visual tasks by combining image-level and block-level contrastive objectives, allowing the model to extract robust, invariant features. Techniques such as masked image modeling (MIM), high-resolution fine-tuning, and the use of regularizers like KoLeo further enhance its ability to capture both global structure and fine-grained details [97]. These capabilities make DINOv2 particularly effective in class-imbalanced medical imaging scenarios, where

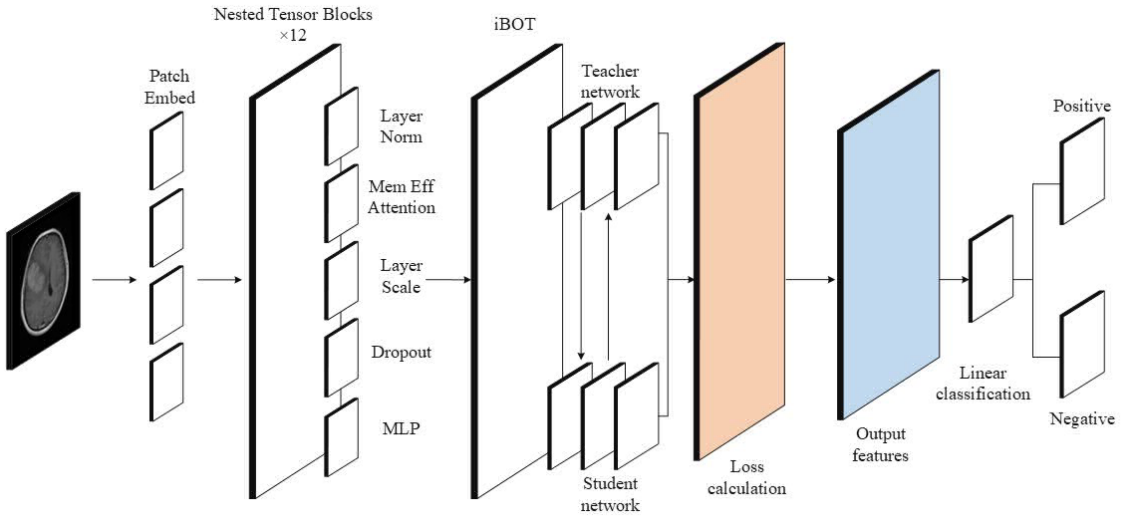


FIGURE 3.1: The DINOv2 framework

labeled data is limited and minority class features are underrepresented [97, 98, 99]. By leveraging clean, diverse pre-training data and avoiding task-specific fine-tuning, DINOv2 provides a flexible and computationally efficient foundation for downstream classification tasks [97, 98, 99].

3.1.2 Self-Calibrated Feature Denoising (SCFD) Method

We introduce a *Self-Calibrated Feature Denoising (SCFD)* framework, consisting of three principal components, as shown in Fig. 3.2: Pre-trained DINOv2 Feature Extractor, Self-Calibrated Auto-encoding for Minority Representation, and Denoised Inference for Minority Classes.

The proposed framework aims to improve the quality and structural consistency of feature embeddings associated with minority classes. These embeddings are extracted by the first component, a *Pre-trained DINOv2 Feature Extractor*, which serves as a fixed backbone model throughout the entire process. Initially, features are obtained for the entire labeled dataset and subsequently partitioned into majority and minority classes based on the class distribution.

In the *Self-Calibrated Auto-encoding for Minority Representation* module, an autoencoder is trained exclusively on minority class features to suppress noise and improve representational quality. The training objective minimizes the mean squared reconstruction error, enabling the model to recover more reliable feature representations. To further constrain the latent space, a self-calibration loss is introduced by aligning the minority latent embeddings with a reference distribution derived from the majority class latent space using Kullback–Leibler divergence. This regularization encourages a more structured and coherent latent representation that reflects the statistical properties of well-formed majority class features.

Following training, the *Denoised Inference for Minority Classes* component applies the trained autoencoder to refine all minority class embeddings. These denoised features are then integrated with the original majority class features to form a unified representation. The resulting feature set is then used to train a downstream classifier, improving classification performance and robustness, particularly for minority classes with limited sample availability.

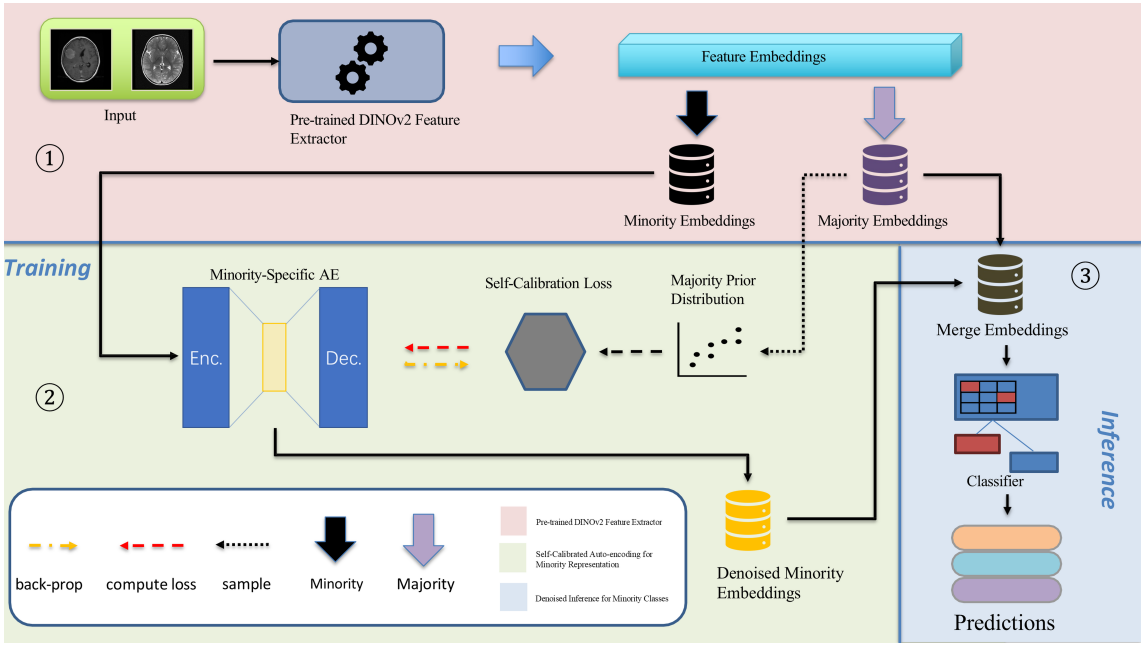


FIGURE 3.2: Our proposed Self-Calibrated Feature Denoising (SCFD) framework

3.1.2.1 Pre-trained DINOv2 Feature Extractor

Our framework begins by extracting feature embeddings from all training samples using a fixed, pretrained DINOv2 model. For each input sample x_i, the DINOv2 encoder $\phi(\cdot)$ generates a feature vector:

$$f_i = \phi(x_i), \quad f_i \in \mathbb{R}^d, \tag{3.1}$$

where d is the feature dimension. Based on the ground-truth class labels, the extracted embeddings are partitioned into two disjoint sets: majority class embeddings $\{f_i^{(maj)}\}$ and minority class embeddings $\{f_i^{(min)}\}$. Due to the scarcity of samples in the minority classes, these embeddings are prone to exhibiting higher noise and variability, motivating targeted denoising.

3.1.2.2 Self-Calibrated Auto-encoding for Minority Representation

In imbalanced datasets, minority class feature embeddings extracted from a pretrained model often exhibit greater noise and reduced structural consistency, primarily due to limited sample availability during representation learning. To address this challenge, we propose a Self-Calibrated Feature Denoising (SCFD) framework that leverages an autoencoder trained exclusively on minority class embeddings to improve their quality, without retraining the large pretrained backbone (we use DINOv2 as the pretrained backbone).

The core component is a lightweight autoencoder (AE), consisting of an encoder $E : \mathbb{R}^d \to \mathbb{R}^{d_z}$ and a decoder $D : \mathbb{R}^{d_z} \to \mathbb{R}^d$. The encoder compresses the input feature embedding f_i into a latent code z_i, and the decoder reconstructs $\hat{f}_i$ from z_i:

$$z_i = E(f_i), \quad \hat{f}_i = D(z_i). \tag{3.2}$$

Training the autoencoder involves minimizing the mean squared reconstruction error over the minority class dataset of size N:

$$\mathcal{L}_{\text{recon}} = \frac{1}{N} \sum_{i=1}^{N} \|f_i - \hat{f}_i\|_2^2, \tag{3.3}$$

which encourages the AE to capture essential structures of minority embeddings while filtering out noise.

To further promote structure in the latent space, we introduce a self-calibration mechanism that regularizes the minority latent representations by aligning them with those of the majority class. Specifically, let $p(z)$ and $q(z)$ denote the latent distributions of majority and minority embeddings, respectively, obtained by encoding the corresponding features:

$$\{z_j^{(maj)} = E(f_j^{(maj)})\}_{j=1}^{N_{maj}}, \quad \{z_i^{(min)} = E(f_i^{(min)})\}_{i=1}^{N}, \tag{3.4}$$

where $f_j^{(maj)}$ and $f_i^{(min)}$ denote majority and minority class embeddings, and N_{maj} is the number of majority samples.

In practice, $p(z)$ and $q(z)$ are approximated as multivariate Gaussian distributions with diagonal covariance matrices. Their means and variances for each latent dimension $k = 1, \ldots, d_z$ are computed empirically as:

$$\mu_p^{(k)} = \frac{1}{N_{maj}} \sum_{j=1}^{N_{maj}} z_j^{(maj,k)}, \quad \sigma_p^{2(k)} = \frac{1}{N_{maj}} \sum_{j=1}^{N_{maj}} \left(z_j^{(maj,k)} - \mu_p^{(k)}\right)^2 + \epsilon, \tag{3.5}$$

$$\mu_q^{(k)} = \frac{1}{N} \sum_{i=1}^{N} z_i^{(min,k)}, \quad \sigma_q^{2(k)} = \frac{1}{N} \sum_{i=1}^{N} \left(z_i^{(min,k)} - \mu_q^{(k)}\right)^2 + \epsilon, \tag{3.6}$$

where ϵ is a small constant for numerical stability.

The self-calibration loss is defined as the Kullback-Leibler (KL) divergence between the two Gaussian distributions:

$$\mathcal{L}_{\text{SCFD}} = D_{\text{KL}}\big(\mathcal{N}(\mu_q, \Sigma_q) \| \mathcal{N}(\mu_p, \Sigma_p)\big), \tag{3.7}$$

where Σ_p and Σ_q are diagonal covariance matrices formed from the variances $\sigma_p^{2(k)}$ and $\sigma_q^{2(k)}$. This KL divergence has a closed-form expression for diagonal Gaussians:

$$D_{\mathrm{KL}} = \frac{1}{2} \sum_{k=1}^{d_z} \left[\log \frac{\sigma_p^{2(k)}}{\sigma_q^{2(k)}} - 1 + \frac{\sigma_q^{2(k)}}{\sigma_p^{2(k)}} + \frac{\left(\mu_p^{(k)} - \mu_q^{(k)}\right)^2}{\sigma_p^{2(k)}} \right]. \tag{3.8}$$

The total loss for training the autoencoder combines reconstruction and self-calibration terms:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{recon}} + \lambda \mathcal{L}_{\mathrm{SCFD}}, \tag{3.9}$$

where the weighting coefficient λ is adaptively computed per training batch as

$$\lambda = \frac{1}{B} \sum_{i=1}^{B} \mathrm{std}(z_i), \tag{3.10}$$

with B denoting batch size, z_i the latent code of the i-th sample, and $\mathrm{std}(\cdot)$ computing the standard deviation over latent features. This dynamic weighting adjusts regularization strength based on latent feature dispersion, alleviating the need for manual hyperparameter tuning.

In summary, this design ensures that minority embeddings are both denoised through reconstruction and regularized to reside in a latent space consistent with majority class features, thereby improving representation robustness and downstream classification performance.

3.1.2.3 Denoised Inference for Minority Classes

Following the training phase, the autoencoder acts as a dedicated denoising module applied exclusively to the minority class feature embeddings. For each minority sample with original embedding $f_i^{(min)} \in \mathbb{R}^d$, the denoised embedding $\tilde{f}_i \in \mathbb{R}^d$ is computed by first encoding into the latent space $\mathbb{R}^{d_z}$ via the encoder E, then reconstructing through the decoder D:

$$\tilde{f}_i = D\big(E(f_i^{(min)})\big). \tag{3.11}$$

The denoising effect stems from the compression inherent in the encoding step,

$$z_i = E(f_i^{(min)}), \quad z_i \in \mathbb{R}^{d_z}, \quad d_z \ll d, \tag{3.12}$$

which forces the latent code z_i to retain only the essential structural components of $f_i^{(min)}$ while discarding noise and minor perturbations.

The decoder D then reconstructs the feature:

$$\tilde{f}_i = D(z_i), \tag{3.13}$$

which approximates the original $f_i^{(min)}$ with noise filtered out.

Let $\mathcal{F}_{min} = \{f_i^{(min)}\}_{i=1}^{N}$ denote the set of minority embeddings, and $\mathcal{F}_{maj} = \{f_j^{(maj)}\}_{j=1}^{N_{maj}}$ the majority embeddings. The final refined feature set $\mathcal{F}_{ref}$ is constructed by merging the denoised minority embeddings with the untouched majority embeddings:

$$\mathcal{F}_{ref} = \mathcal{F}_{maj} \cup \{\tilde{f}_i \mid f_i^{(min)} \in \mathcal{F}_{min}\}. \tag{3.14}$$

This selective denoising preserves the intrinsic distribution of the majority features while improving the quality and structural consistency of the minority features.

The refined feature set $\mathcal{F}_{ref}$ is subsequently used to train downstream classifiers $\mathcal{C}$, which learn decision boundaries in the improved embedding space:

$$\hat{y} = \mathcal{C}(\mathcal{F}_{ref}), \tag{3.15}$$

where $\hat{y}$ denotes predicted class labels.

By leveraging the autoencoder for denoising without retraining the large pretrained backbone (in our framework, instantiated by DINOv2), the SCFD framework effectively enhances class separability and robustness. This approach mitigates the noise and variability inherent in minority class data, leading to improved classification performance under severe class imbalance.

3.2 Experiment

3.2.1 Datasets

We utilized three datasets in this study. The distribution of these datasets is summarized in Table 3.1.

TABLE 3.1: Summary of datasets used in our experiments.

Dataset	Exam.	Labels	Classes	Images
Lab-shanghai	MRI	Cerebral infarction	2	23,148
Brain Tumor 4C [100]	MRI	Brain tumor	4	7,023
Chest CT [101]	CT	Lung tumor	4	1,000

The Lab-shanghai dataset consists of cerebral infarction data obtained from the Institute of Brain Sciences, Shanghai University of Health. 11,574 cerebral infarction cases and 11,574 normal cases, with 2,895 patients aged 28 to 88 years, from 2020-2022, were included in the study. There were 1450 male patients and 1445 female patients. DWI images were obtained for all participants. Image labeling was performed by 2 radiologists with more than 20 years of experience. The purpose of this data collection is to classify and predict patients based on their MRI. The cerebral infarction category is labeled as "yes," and the normal picture is "no". Informed consent was obtained from the patients by the Institute of Brain Sciences, and ethical approval for the use of secondary data was granted by UTS under approval number ETH24-10033. The Brain Tumor 4C dataset [100] consists of MRI scans annotated for brain tumor classification. It contains 7,023 labeled images spanning 4 classes. The Chest CT dataset [101] includes 1,000 CT scans focused on lung tumor identification. It features 4 classes with detailed lesion annotations, facilitating the study of pulmonary abnormalities.

Both datasets are sourced from the Kaggle website [100, 101], which provides curated collections of medical images to support research in computer-aided diagnosis. They complement our experiments by covering different imaging modalities (MRI and CT) and tumor types (brain and lung), thereby demonstrating the generalizability of our proposed framework across diverse clinical classification scenarios. The dataset was split into training, validation, and test sets in a 70:10:20 ratio using stratified sampling to preserve class

distributions, with the smaller validation set used for hyperparameter tuning and the test set for final evaluation.

3.2.2 Imbalance Settings

In our experiments, the number of samples in the normal class (majority class) was kept unchanged, while the number of samples in the lesion class (minority class) was reduced to simulate different degrees of class imbalance. Specifically, the lesion class sample size was sequentially reduced to represent extreme imbalance (5%), high imbalance (20%), moderate imbalance (50%), and the original distribution (100%). For example, 5% indicates that the number of lesion class samples is only 5% of the original count, while the number of normal class samples remains intact. By artificially reducing the minority class samples, we create training sets with varying imbalance ratios to evaluate model performance across different class-imbalance scenarios.

3.2.3 Compared Metrics

We evaluate model performance using the macro average of Recall, Precision, and F1 score. Given C classes, these metrics are computed as follows:

$$\text{Recall}_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FN_c} \tag{3.16}$$

$$\text{Precision}_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FP_c} \tag{3.17}$$

$$F1_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C}\frac{2 \times \text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{3.18}$$

where TP_c, FP_c, and FN_c denote true positives, false positives, and false negatives for class c, respectively.

Comparisons among supervised models, the self-supervised DINOv2, and our proposed SCFD framework show that SCFD consistently outperforms the others, particularly under severe class imbalance.

3.2.4 Implementation Details

All experiments were conducted on a computer equipped with an NVIDIA GeForce RTX 4090 GPU, an Intel Core i9-13900KF CPU, and 64 GB of DDR5 RAM operating at 5200 MHz. The entire training pipeline was executed on Windows 11. The self-supervised model (DINOv2) was fine-tuned for 50 epochs with a learning rate of 1e-6 and a batch size of 16. The training process employed the Cross-Entropy Loss and was optimized using the Adam optimizer. Our method, SCFD, was trained for 50 epochs using the Adam optimizer with a learning rate of 1e-3, decaying by 1% every epoch, and a batch size of 16. The training objective combined a reconstruction loss (mean squared error) and a self-calibration loss.

Popular supervised models such as ResNet50, VGG16, and DenseNet121 were trained under standardized settings for fair comparison. The loss function used was Cross-Entropy Loss, which is widely adopted to measure the discrepancy between predicted and ground-truth labels. For optimization, we employed Stochastic Gradient Descent (SGD) with a momentum of 0.9 to enhance convergence speed, and the initial learning rate was set to 0.001. A cosine annealing learning rate scheduler was applied to adjust the learning rate during training. All models were trained for 50 epochs with a batch size of 16.

3.2.5 Computational Complexity Analysis

The SCFD introduces negligible computational overhead as it operates solely on the extracted low-dimensional feature embeddings. Let N_m denote the number of minority samples, d the original feature dimension, and $d_z \ll d$ the bottleneck dimension of the autoencoder. The computational complexity of training the autoencoder is $O(N_m \cdot d \cdot d_z \cdot E)$, where E is the number of training epochs. Estimating the prior distribution from majority class embeddings involves a complexity of $O(N_M \cdot d_z^2)$, where $N_M \ll N$ denotes the

number of majority samples selected for prior modeling. The KL divergence computation required for self-calibration introduces only $O(d_z)$ complexity per sample. At inference time, the overhead introduced by the autoencoder for each minority sample is merely $O(d \cdot d_z)$, making SCFD computationally lightweight and easily integrable into existing feature-based pipelines.

TABLE 3.2: Classification performance for Lab-shanghai dataset of various models evaluated at different numbers of lesions (100%, 50%, 20%, and 5%). The table reports Macro-avg Precision, Recall, and F1-score to assess model effectiveness across data scales.

Model	100%			50%			20%			5%		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
ResNet18	0.967	0.967	0.967	0.898	0.890	0.894	0.911	0.769	0.822	0.991	0.643	0.718
ResNet34	0.967	0.967	0.967	0.911	0.894	0.902	0.933	0.725	0.791	0.991	0.643	0.718
ResNet50	0.963	0.985	0.973	0.915	0.920	0.917	0.982	0.773	0.844	0.488	0.500	0.494
ResNet101	0.963	0.985	0.973	0.909	0.918	0.914	0.974	0.659	0.728	0.991	0.643	0.718
ResNet152	0.963	0.985	0.973	0.936	0.913	0.924	0.970	0.614	0.670	0.993	0.714	0.796
VGG11	0.964	0.959	0.962	0.934	0.956	0.945	0.799	0.544	0.561	0.488	0.500	0.494
VGG13	0.974	0.983	0.978	0.950	0.955	0.952	0.886	0.701	0.759	0.488	0.500	0.494
VGG16	0.975	0.970	0.973	0.966	0.946	0.955	0.863	0.722	0.772	0.488	0.500	0.494
VGG19	0.994	0.984	0.989	0.956	0.956	0.956	0.952	0.816	0.870	0.993	0.714	0.796
MobileNetV2	0.881	0.897	0.889	0.364	0.500	0.421	0.462	0.500	0.480	0.488	0.500	0.494
DenseNet121	0.977	0.991	0.984	0.943	0.934	0.938	0.890	0.790	0.831	0.993	0.714	0.796
DenseNet161	0.977	0.991	0.984	0.954	0.949	0.952	0.984	0.795	0.863	0.995	0.786	0.861
DenseNet169	0.970	0.988	0.979	0.947	0.929	0.938	0.939	0.748	0.813	0.993	0.714	0.796
DenseNet201	0.970	0.988	0.979	0.947	0.929	0.938	0.956	0.839	0.887	0.993	0.714	0.796
ShuffleNetV2 ×0.5	0.963	0.985	0.973	0.925	0.916	0.921	0.923	0.814	0.859	0.991	0.643	0.718
ShuffleNetV2 ×1.0	0.977	0.991	0.984	0.950	0.936	0.942	0.948	0.794	0.852	0.895	0.784	0.830
ShuffleNetV2 ×1.5	0.970	0.988	0.979	0.941	0.927	0.933	0.917	0.792	0.841	0.993	0.714	0.796
ShuffleNetV2 ×2.0	0.963	0.985	0.973	0.907	0.911	0.909	0.987	0.841	0.899	0.993	0.714	0.796
MnasNet 0.5	0.364	0.500	0.421	0.398	0.500	0.443	0.462	0.500	0.480	0.488	0.500	0.494
MnasNet 1.0	0.364	0.500	0.421	0.398	0.500	0.443	0.462	0.500	0.480	0.488	0.500	0.494
ViT-B/16	0.985	0.986	0.985	0.747	0.747	0.747	0.637	0.614	0.624	0.488	0.500	0.494
DINOv2	**0.993**	0.992	0.992	**0.992**	0.990	0.991	**0.994**	0.973	0.983	**0.992**	0.959	0.975
SCFD	0.991	**0.995**	**0.993**	**0.992**	**0.992**	**0.992**	0.985	**0.986**	**0.986**	0.983	**0.976**	**0.979**

3.3 Results

3.3.1 Results on Lab-shanghai Dataset

Table 3.2 summarizes the classification performance of various models on the Lab-shanghai dataset under different lesion (minority) class ratios: 100%, 50%, 20%, and 5%, with the number of normal (majority) class samples kept constant. Macro-avg Precision, Recall, and F1-score are reported to evaluate model effectiveness across varying levels of class imbalance. As the proportion of lesion samples decreases, conventional supervised convolutional models exhibit a significant drop in F1-score, reflecting their sensitivity to class

TABLE 3.3: Classification performance for the Brain Tumor 4C dataset of various models evaluated at different numbers of lesions (100%, 50%, 20%, and 5%). The table reports Macro-avg Precision, Recall, and F1-score to assess model effectiveness across data scales.

Model	100%			50%			20%			5%		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
ResNet18	0.944	0.943	0.943	0.878	0.873	0.874	0.810	0.803	0.802	0.709	0.700	0.690
ResNet34	0.941	0.940	0.941	0.864	0.860	0.860	0.783	0.771	0.773	0.719	0.706	0.697
ResNet50	0.940	0.939	0.940	0.844	0.842	0.841	0.782	0.774	0.773	0.676	0.674	0.661
ResNet101	0.929	0.929	0.929	0.879	0.876	0.876	0.799	0.791	0.790	0.672	0.667	0.654
ResNet152	0.936	0.935	0.935	0.864	0.860	0.861	0.791	0.785	0.784	0.639	0.639	0.623
VGG11	0.956	0.954	0.954	0.900	0.894	0.895	0.804	0.796	0.796	0.684	0.658	0.652
VGG13	0.957	0.956	0.957	0.899	0.893	0.894	0.804	0.797	0.797	0.712	0.699	0.689
VGG16	0.956	0.954	0.955	0.903	0.898	0.899	0.816	0.810	0.809	0.711	0.698	0.690
VGG19	0.960	0.959	0.960	0.907	0.901	0.902	0.800	0.796	0.794	0.689	0.681	0.672
MobileNetV2	0.933	0.931	0.932	0.777	0.771	0.767	0.077	0.250	0.118	0.077	0.250	0.118
DenseNet121	0.954	0.953	0.954	0.909	0.905	0.906	0.825	0.816	0.818	0.734	0.726	0.718
DenseNet161	0.961	0.961	0.961	0.910	0.905	0.906	0.841	0.830	0.832	0.733	0.726	0.718
DenseNet169	0.953	0.952	0.952	0.904	0.900	0.901	0.828	0.819	0.820	0.711	0.706	0.695
DenseNet201	0.959	0.959	0.959	0.903	0.899	0.900	0.838	0.829	0.831	0.715	0.709	0.697
ShuffleNetV2 ×0.5	0.904	0.903	0.903	0.841	0.840	0.840	0.703	0.700	0.697	0.157	0.306	0.200
ShuffleNetV2 ×1.0	0.924	0.922	0.922	0.844	0.835	0.837	0.688	0.656	0.660	0.409	0.520	0.451
ShuffleNetV2 ×1.5	0.915	0.914	0.914	0.837	0.828	0.831	0.813	0.804	0.805	0.394	0.491	0.424
ShuffleNetV2 ×2.0	0.903	0.901	0.902	0.852	0.846	0.848	0.749	0.740	0.740	0.377	0.426	0.341
MnasNet 0.5	0.077	0.250	0.118	0.077	0.250	0.118	0.077	0.250	0.118	0.077	0.250	0.118
MnasNet 1.0	0.077	0.250	0.118	0.077	0.250	0.118	0.077	0.250	0.118	0.077	0.250	0.118
ViT-B/16	0.859	0.859	0.859	0.810	0.808	0.808	0.735	0.723	0.723	0.635	0.576	0.562
DINOv2	**0.996**	**0.996**	**0.996**	0.978	0.976	0.977	0.961	0.957	0.958	0.857	0.837	0.842
SCFD	0.993	0.993	0.993	**0.990**	**0.989**	**0.989**	**0.965**	**0.963**	**0.963**	**0.893**	**0.887**	**0.888**

imbalance. For instance, ResNet50's F1-score declines sharply from 0.973 under balanced conditions to 0.494 at 5% minority samples. Similarly, VGG variants (VGG11, VGG13, VGG16) and MnasNet suffer notable performance degradation at this extreme imbalance, with F1-scores falling below 0.5. The ViT-B/16 model also shows a marked decrease to 0.494. In contrast, DenseNet and ShuffleNetV2 demonstrate relatively better robustness, though they are still affected.

In contrast, the self-supervised DINOv2 model exhibits markedly higher and more stable performance across all data scales. It achieves an F1-score of 0.992 under balanced settings and maintains an impressive 0.975 even under extreme imbalance (5% lesion data). DINOv2 consistently surpasses all supervised models in Precision, Recall, and F1-score, demonstrating strong robustness to data scarcity. Our proposed Self-Calibrated Feature Denoising (SCFD) framework further addresses class imbalance by enhancing representations of minority classes. Under balanced conditions, SCFD performs comparably to DINOv2, achieving slightly higher Recall and F1-score (0.995 and 0.993 vs. 0.992 and 0.992), with a marginal trade-off in Precision (0.991 vs. 0.993). When the lesion data is

TABLE 3.4: Classification performance for Chest CT dataset of various models evaluated at different numbers of lesions (100%, 50%, 20%, and 5%). The table reports Macro-avg Precision, Recall, and F1-score to assess model effectiveness across data scales.

Model	100%			50%			20%			5%		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
ResNet18	0.554	0.563	0.518	0.548	0.536	0.514	0.456	0.508	0.443	0.325	0.490	0.384
ResNet34	0.619	0.610	0.587	0.614	0.584	0.554	0.450	0.506	0.450	0.435	0.478	0.441
ResNet50	0.639	0.585	0.526	0.518	0.552	0.477	0.455	0.501	0.429	0.322	0.488	0.381
ResNet101	0.571	0.592	0.564	0.470	0.508	0.463	0.506	0.500	0.472	0.440	0.483	0.418
ResNet152	0.514	0.534	0.484	0.531	0.535	0.507	0.365	0.500	0.407	0.332	0.500	0.376
vgg11	0.583	0.620	0.536	0.533	0.595	0.538	0.236	0.444	0.307	0.043	0.250	0.073
vgg13	0.597	0.650	0.569	0.548	0.549	0.476	0.182	0.371	0.229	0.108	0.260	0.095
vgg16	0.623	0.635	0.541	0.495	0.532	0.466	0.167	0.342	0.200	0.043	0.250	0.073
vgg19	0.627	0.626	0.534	0.499	0.540	0.478	0.144	0.306	0.161	0.117	0.263	0.099
MobileNetV2	0.095	0.250	0.138	0.043	0.250	0.073	0.043	0.250	0.073	0.043	0.250	0.073
DenseNet121	0.549	0.582	0.507	0.533	0.549	0.514	0.584	0.512	0.498	0.568	0.493	0.440
DenseNet161	0.584	0.571	0.497	0.488	0.516	0.481	0.508	0.515	0.497	0.435	0.501	0.427
DenseNet169	0.542	0.572	0.524	0.496	0.523	0.487	0.524	0.507	0.481	0.389	0.488	0.409
DenseNet201	0.550	0.569	0.513	0.512	0.540	0.495	0.667	0.465	0.445	0.205	0.392	0.249
ShuffleNetV2 ×0.5	0.095	0.250	0.138	0.043	0.250	0.073	0.043	0.250	0.073	0.043	0.250	0.073
ShuffleNetV2 ×1.0	0.095	0.250	0.138	0.043	0.250	0.073	0.043	0.250	0.073	0.043	0.250	0.073
ShuffleNetV2 ×1.5	0.095	0.250	0.138	0.043	0.250	0.073	0.043	0.250	0.073	0.043	0.250	0.073
ShuffleNetV2 ×2.0	0.095	0.250	0.138	0.043	0.250	0.073	0.043	0.250	0.073	0.043	0.250	0.073
MnasNet 0.5	0.095	0.250	0.138	0.043	0.250	0.073	0.043	0.250	0.073	0.043	0.250	0.073
MnasNet 1.0	0.095	0.250	0.138	0.043	0.250	0.073	0.043	0.250	0.073	0.043	0.250	0.073
ViT-B/16	0.490	0.519	0.473	0.476	0.501	0.463	0.456	0.497	0.445	0.443	0.493	0.437
DINOv2	**0.915**	**0.933**	**0.919**	0.830	0.832	0.830	**0.748**	0.730	0.731	0.665	0.627	0.583
SCFD	0.887	0.867	0.874	**0.853**	**0.853**	**0.851**	0.744	**0.786**	**0.745**	**0.685**	**0.728**	**0.669**

reduced to 50%, both models perform similarly. However, under severe imbalance (20% and 5% lesion data), SCFD consistently outperforms DINOv2 in Recall and F1-score, which are critical for accurate minority class detection. For instance, at 5% lesion data, SCFD achieves a Recall of 0.976 and an F1-score of 0.979, compared to DINOv2's 0.959 and 0.975. Precision remains comparable, with only marginal differences. These results highlight the limitations of traditional supervised models under data-scarce, imbalanced conditions. While DINOv2 demonstrates strong self-supervised learning capabilities, its performance degrades under extreme imbalance. SCFD mitigates this issue by reinforcing minority class features, improving recall and maintaining balanced performance under challenging class-imbalance scenarios.

3.3.2 Results on Brain Tumor 4C Dataset

Table 3.3 presents the classification performance of various models on the Brain Tumor 4C dataset under simulated imbalance conditions, where the lesion (minority) class was reduced to 100%, 50%, 20%, and 5% of its original count, while keeping the number of

normal (majority) class samples fixed. Macro-avg Precision, Recall, and F1-score are reported to assess model performance under varying levels of label imbalance.

As the imbalance becomes more severe, F1-scores of most supervised convolutional networks drop significantly, highlighting their strong reliance on labeled lesion samples. For instance, ResNet18 sees its F1-score drop from 0.943 with 100% lesion data to just 0.690 with 5%. Similar patterns are observed for deeper ResNet variants and VGG networks, all of which experience marked performance degradation as the amount of lesion data decreases. Notably, lightweight architectures such as MobileNetV2, ShuffleNetV2 ($\times$0.5–2.0), and especially MnasNet (both 0.5 and 1.0) perform poorly under extreme imbalance (5%), with F1-scores stagnating around 0.118 or lower, indicating failure to learn meaningful representations for the minority class.

The self-supervised model DINOv2 consistently outperforms supervised models across all sample sizes. Under balanced conditions (100% lesion data), it slightly surpasses SCFD in all metrics, achieving an F1-score of 0.996 versus 0.993. This suggests that DINOv2 is particularly effective when sufficient labeled data is available. However, as lesion data becomes increasingly limited (20% and 5%), SCFD demonstrates clear advantages. It outperforms DINOv2 in Recall and F1-score, two metrics that are especially important for identifying tumor categories in class imbalance settings. At 5% lesion data, SCFD achieves an F1-score of 0.888, substantially higher than DINOv2's 0.842, indicating a better effect on class imbalance. In summary, while DINOv2 leverages strong self-supervised representations under full supervision, SCFD offers greater robustness in imbalanced and limited-data settings. These results validate the proposed method's effectiveness in real-world scenarios where annotated lesion data is sparse.

3.3.3 Results on Chest CT Dataset

Table 3.4 reports the classification results on the Chest CT dataset, where lesion class samples were progressively reduced to 100%, 50%, 20%, and 5% of their original size, simulating varying levels of class imbalance. The number of normal class samples was kept constant. The evaluation metrics include Macro-avg Precision, Recall, and F1-score.

Traditional supervised models such as ResNet, VGG, DenseNet, MobileNet, ShuffleNet, MnasNet, and ViT-B/16 exhibit substantial performance degradation as class imbalance increases. While some of these models show moderate F1-scores in balanced settings (e.g., ResNet34 achieves 0.587 at 100%), their performance drops significantly under extreme imbalance. For instance, ResNet18 and ResNet50 see their F1-scores fall to 0.384 and 0.381, respectively, at the 5% lesion level. VGG variants are particularly affected, with F1-scores declining to nearly random levels (e.g., VGG11 and VGG16 both at 0.073), indicating a complete collapse in minority class prediction.

Lightweight architectures such as MobileNetV2, ShuffleNetV2 ($\times$0.5–2.0), and MnasNet (0.5 and 1.0) consistently yield F1-scores of only 0.073 or lower across all imbalance levels, failing to distinguish lesion cases even when moderately balanced.

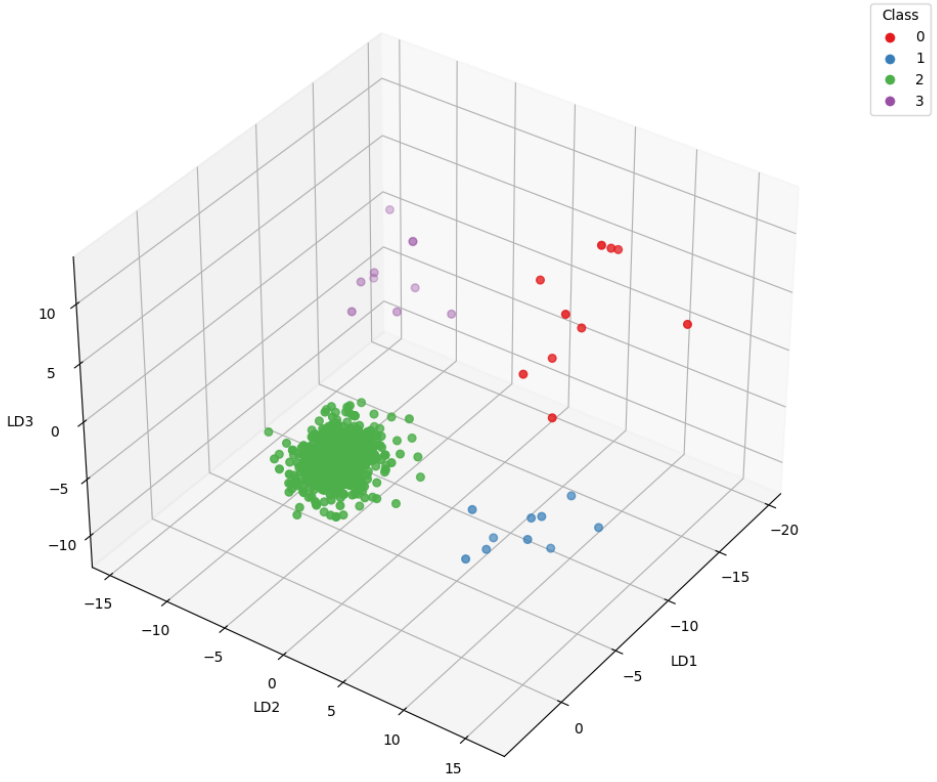
The self-supervised model DINOv2 consistently outperforms all supervised models across all levels of data availability. Under the balanced setting (100% lesion data), it also achieves slightly higher metrics than SCFD (e.g., F1-score of 0.919 vs. 0.874), indicating superior representation capabilities in fully supervised scenarios. Nonetheless, as class imbalance becomes more severe (20% and 5% lesion data), SCFD overtakes DINOv2, particularly in Recall and F1-score. With only 5% of lesion data, SCFD achieves a Recall of 0.728 and an F1-score of 0.669, outperforming DINOv2's corresponding values of 0.627 and 0.583. These findings highlight the limitations of supervised models in class-imbalanced medical scenarios and underscore the strengths of self-supervised learning. Although DINOv2 performs well under class imbalance, the SCFD framework offers greater robustness and reliability in more challenging settings with limited imbalanced data.

3.3.4 Qualitative Visualization Results

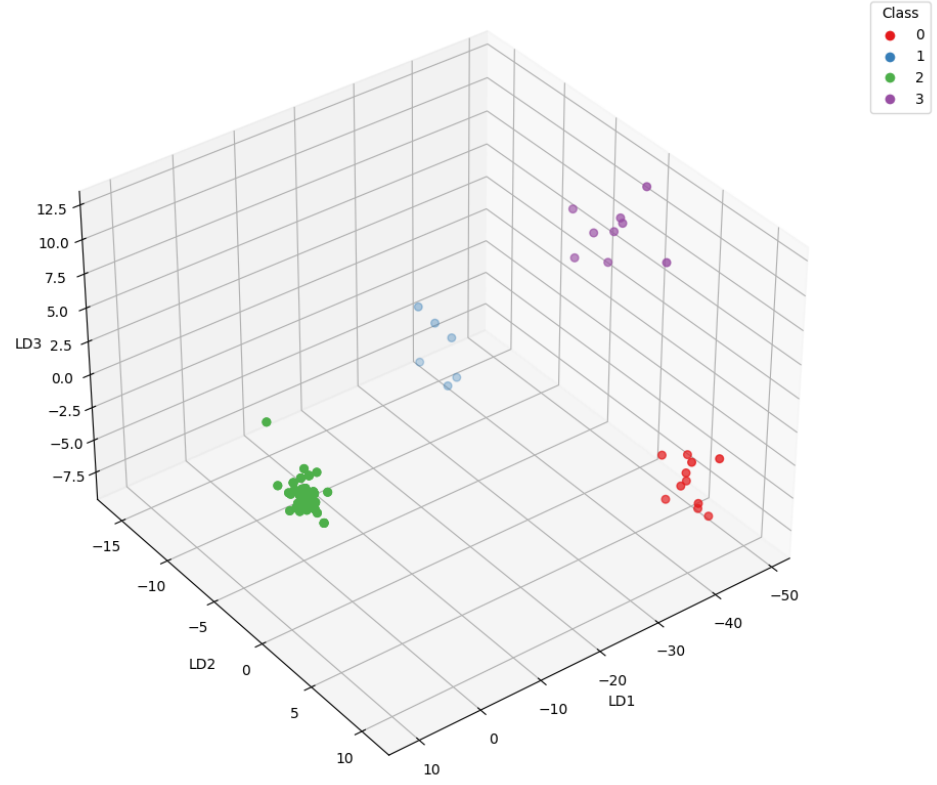
The goal of this visualization is to assess the model's ability to learn discriminative feature representations under severe class imbalance, where the minority class is substantially underrepresented. Across all datasets, the number of normal (majority) samples remains unchanged, while only 5% of lesion (minority) samples are retained to simulate the imbalance. The extracted features are grouped by their ground-truth labels.

Dimensionality reduction is then performed in two steps: Principal Component Analysis (PCA) is applied to retain at least 80% of the variance, followed by Linear Discriminant Analysis (LDA) to project the data into a 3D space that maximizes class separability. Fig. 3.3 shows the resulting embeddings, with classes distinguished by color, demonstrating SCFD's effectiveness in preserving minority class separability despite the imbalance.

In Fig. 3.3a, corresponding to the Brain Tumor 4C dataset with a substantially larger sample size, feature embeddings exhibit clear and well-defined clusters. Each class forms distinct, compact groups with minimal overlap, indicating strong inter-class separability. Fig. 3.3b shows feature embeddings from the Chest CT dataset, characterized by limited samples and severe imbalance. Despite these constraints, the projected samples form four well-separated and cohesive clusters, illustrating SCFD's capacity to learn discriminative representations under challenging data conditions. Collectively, these visualizations provide empirical evidence of SCFD's robustness in preserving class separability and generating structurally coherent embeddings under extreme class imbalance and limited data, critical factors for improving medical image classification performance.



(a) 3D visualization of Brain Tumor 4C dataset.The green part represents the normal category, while the other colors represent the pathological categories.



(b) 3D visualization of Chest CT dataset. The green part represents the normal category, while the other colors represent the pathological categories.

FIGURE 3.3: 3D Visualization of Feature Embeddings via LDA with 5% Lesion Samples and Unchanged Normal Samples.

Chapter 4

Label Uncertainty Transformation for Combating Noisy Labels in Medical Image Segmentation

This chapter introduces the Label Uncertainty Transformation (LUT) framework, developed to address the challenges of noisy labels in medical image segmentation. Due to annotation scarcity and inherent noise, segmentation tasks in medical imaging require methods that can handle both uncertainty at the instance level and ambiguity in target structures. LUT combines uncertainty estimation, entropy-based sample selection, and instance-dependent label transformation into a cohesive and robust training pipeline. Unlike traditional approaches that rely on clean labels or accurate pseudo-labels, LUT leverages deep ensemble techniques and pixel-level entropy information to improve model performance under noisy conditions. The following sections provide a detailed description of the key components—Uncertainty Estimation Ensemble Module, Label Transformation Network (LTN), and Target Segmentation Network (TSN), as well as the training strategy and overall workflow.

TABLE 4.1: Description of symbols

Symbol	Description
μ_1, μ_2	Probability maps from U-Net and nnU-Net
μ_{mean}	Weighted average prediction
σ_1^2, σ_2^2	Variances from deviation to μ_{mean}
w_1, w_2	Weights derived from Dice scores of U-Net and nnU-Net
w_1', w_2'	Inverse-variance weights
$\hat{\mu}$	Final ensemble prediction
ψ^2	Ensemble variance (uncertainty)
$\tilde{Y}$	Noisy labels
$D^*, \tilde{D}$	Distilled subset and noisy dataset
H_i	Pixel entropy at pixel i
$p_{i,c}$	Probability of pixel i in class c
$T(x;\theta)$	Transition matrix from clean to noisy labels
$\bar{y}^*, \tilde{y}^*$	Clean-like prediction and noisy label
L_{rbce}	Rectified BCE loss
$\psi_1(x)$	Combined uncertainty measure
L_{ua}	Uncertainty-weighted loss
L_{entropy}	Entropy regularizer
γ	Entropy regularization weight
L_g	Overall LTN loss
$f(x;\phi)$	Target segmentation network
$\bar{y}, \hat{y}$	Clean prediction and noisy mapping
L_{BCE}	BCE loss for TSN

4.1 Methodology

We propose a statistically consistent prediction model, the Label Uncertainty Transformation (LUT) framework, by parametrically estimating an IDTM for each predicted segmentation mask. While the core idea of using an instance-dependent transition matrix is inspired by Zhang et al.[102], we introduce the following novel strategies specifically tailored for medical image segmentation with highly noisy labels: (1) a *deep ensemble approach* in our Uncertainty Estimation Ensemble Module to handle complex medical data and not relying on a single model, (2) a *pixel-entropy-based sample selection* method to refine distilled examples for robust LTN training, and (3) a *pixel entropy regularizer* to handle sparse target structures common in lesions or tumors. These innovations aim to improve both *accuracy* and *stability* under noisy medical labels.

The LUT consists of three components as shown in Fig. 4.1: the Uncertainty Estimation Ensemble Module, which generates initial segmentation and uncertainty estimates; the Label Transformation Network (LTN), which refines segmentation via an Instance-Dependent Transition Matrix (IDTM); and the Target Segmentation Network (TSN), which optimizes the final output. A reweighting technique adjusts the influence of predictions based on uncertainty, and an entropy regularizer reduces complexity and mitigates

overfitting. This architecture improves segmentation accuracy and reliability, making it well-suited to precision-critical medical image analysis tasks. The complete workflow of LUT is summarized in Algorithm 4.1, effectively integrating uncertainty, entropy-based sample selection, and forward correction to systematically address label noise.

Algorithm 4.1 Label Uncertainty Transformation (LUT)

Noisy set $\tilde{\mathcal{D}} = \{(x, \tilde{y})\}_{n=1}^{N}$ Trained TSN $f(\cdot; \phi)$ for deployment
Step 1: Train UEEM
- Train U-Net and nnU-Net with Eq.(4.9); - Obtain ensemble mean $\hat{\mu}$, variance ψ.
Step 2: Distill Reliable Samples
- Compute pixel entropy in each model; merge results.
- Retain pixels exceeding a global threshold, forming $\mathcal{D}^*$.
Step 3: Learn LTN
- Parameterize $T(x; \theta)$, optimize L_g [Eq.(4.14)] with uncertainty re-weighting, entropy regularization.
- Generate IDTM for each sample.
Step 4: Train TSN
- Fix $T(x; \theta)$; let $\bar{y} = f(x; \phi)$, $\hat{y} = T^{\top}\bar{y}$.
- Minimize $L_{\mathrm{BCE}}(\hat{y}, \tilde{y})$ [Eq.(4.15)].
$f(\cdot; \phi)$.

4.1.1 Preliminaries

In this work, we let $\tilde{\mathcal{D}} = \{(x_n, \tilde{y}_n)\}_{n=1}^{N}$ denote our noisy dataset, where each x_n is an input (e.g., an MRI slice) and $\tilde{y}_n$ is the associated noisy segmentation mask. A subset $D^* \subset \tilde{\mathcal{D}}$ consists of "distilled" samples that have been deemed more reliable, typically via a pixel-entropy screening process. We use ϕ to represent the trainable parameters of our neural networks; for example, ϕ might parameterize the Target Segmentation Network (TSN). Whenever a superscript * appears (e.g., $\bar{y}^*$), it indicates either a distilled variant of a variable (as with sample x^* in D^*) or a refined model prediction believed to be closer to the clean label distribution. Additionally, we define $T(x)$ to be an instance-dependent transition matrix (IDTM) for input image x, capturing the probability of label flips from a true label e^i to a noisy label e^j. When the TSN produces a "clean" prediction $\bar{y}$, we obtain the corresponding noisy-space prediction $\hat{y}$ by multiplying $\bar{y}$ with $T(x)^{\top}$. This forward-correction step aligns the learned clean labels with the observed noisy labels during training. We also employ μ, σ^2 to denote the mean and variance generated by our deep ensemble's uncertainty estimation, and define ψ as an overall predictive variance

measure. Finally, our training objectives make use of several specialized loss functions, including L_{BCE} (binary cross-entropy), L_{rbce} (rectified BCE), L_{ua} (an uncertainty-adjusted variant), and L_{entropy} (a pixel-level entropy regularizer).

4.1.2 IDTM Formulation and Noisy Label Modeling

In medical image segmentation, manual annotations often contain inaccuracies. We denote by $\tilde{Y}$ the noisy labels, Y the true labels, and X the input image (e.g., an MRI slice). Label noise may depend on local image characteristics—for instance, pixels near lesion boundaries or low-contrast regions are more vulnerable to mislabeling.

To capture such instance-dependent noise, we introduce a matrix $T(x) \in [0,1]^{2\times 2}$ for each image x. Each entry $T_{i,j}(x)$ indicates the probability of a true label e^i being corrupted into the noisy label e^j. Formally, if we let

$$p_i(x) = \mathbb{P}(Y = e^i \mid X = x), \tag{4.1}$$

$$\tilde{p}_j(x) = \mathbb{P}(\tilde{Y} = e^j \mid X = x), \tag{4.2}$$

then $T(x)$ provides a mapping from the clean to the noisy distribution as follows:

$$\tilde{p}_j(x) \;=\; \sum_i T_{i,j}(x)\, p_i(x). \tag{4.3}$$

By appropriately estimating $T(x)$, we can disentangle the noisy labels from their underlying clean counterparts and perform a forward correction step to guide the training process. This framework is particularly crucial in medical imaging scenarios where noise is often structured by anatomical or acquisition-related factors (e.g., boundary ambiguity, partial volume effects).

4.1.3 Uncertainty Estimation Ensemble Module (UEEM)

Our framework begins by independently training two segmentation networks (U-Net and nnU-Net) on the noisy dataset. Each model produces a lesion probability map, denoted as

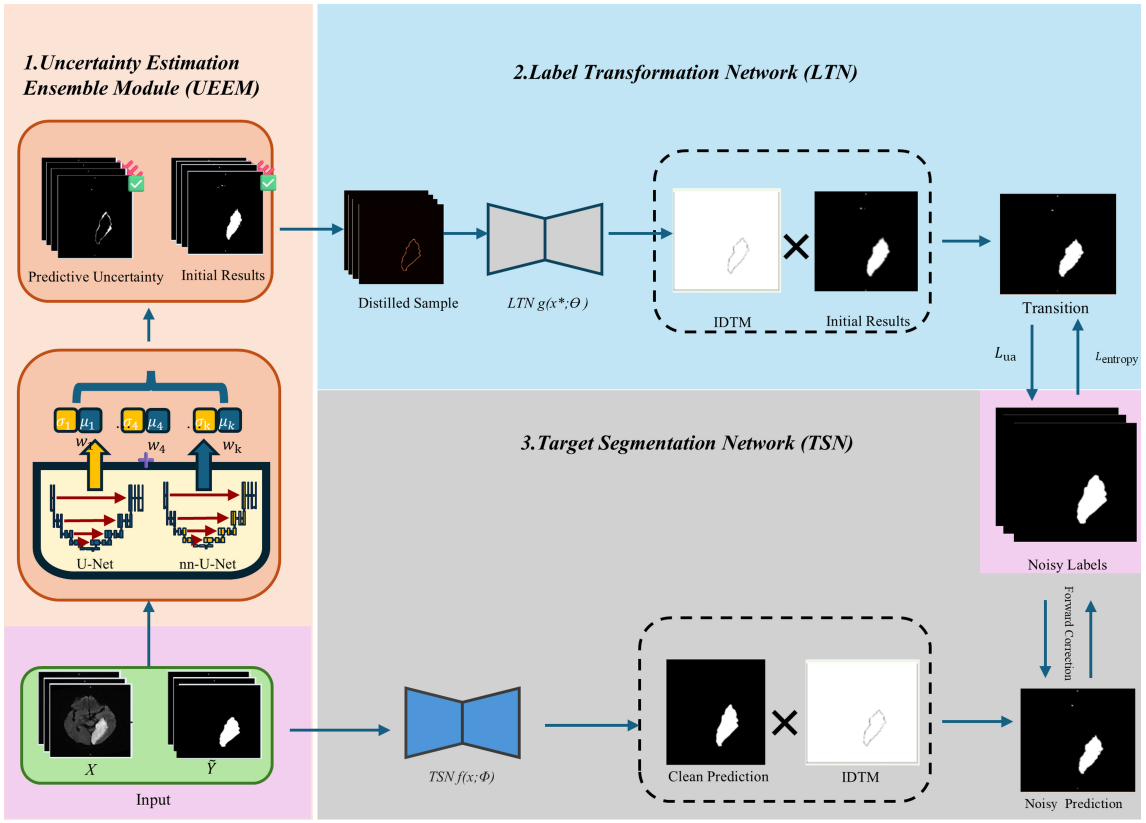


FIGURE 4.1: Our proposed Label Uncertainty Transformation (LUT) framework

μ_1 and μ_2, respectively. Instead of assuming each model provides a predictive variance, we adopt a deep ensemble strategy that estimates uncertainty from the discrepancy between model predictions.

The ensemble mean prediction is computed by weighting the two models' outputs according to their Dice scores on a validation set:

$$\mu_{\text{mean}} = \frac{w_1\mu_1 + w_2\mu_2}{w_1 + w_2} \tag{4.4}$$

where w_1 and w_2 represent the Dice scores of the U-Net and nnU-Net models, respectively.

To approximate the confidence of each model's prediction, we estimate its predictive variance using its deviation from the ensemble mean:

$$\sigma_1^2 = (\mu_1 - \mu_{\text{mean}})^2, \quad \sigma_2^2 = (\mu_2 - \mu_{\text{mean}})^2 \tag{4.5}$$

The inverse-variance weights are then computed as:

$$w_1' = \frac{1}{\sigma_1^2}, \quad w_2' = \frac{1}{\sigma_2^2} \tag{4.6}$$

Using these weights, the final ensemble prediction is refined via precision-weighted averaging:

$$\hat{\mu} = \frac{w_1'\mu_1 + w_2'\mu_2}{w_1' + w_2'} \tag{4.7}$$

The ensemble variance ψ^2, which captures epistemic uncertainty from model disagreement, is calculated as:

$$\psi^2 = \frac{w_1'(\mu_1 - \hat{\mu})^2 + w_2'(\mu_2 - \hat{\mu})^2}{w_1' + w_2'} \tag{4.8}$$

To regularize training under noisy supervision, we define the loss as the sum of binary cross-entropy (BCE) and the ensemble variance:

$$\mathcal{L}_{\text{UEEM}} = \text{BCE}(\tilde{Y}, \hat{\mu}) + \psi^2 \tag{4.9}$$

where $\tilde{Y}$ denotes the noisy labels. This formulation allows the model to naturally suppress uncertain predictions while improving the reliability of the learned representation.

After convergence, the final prediction $\hat{\mu}(x)$ and the uncertainty map $\psi(x)$ are obtained for each pixel x. High-uncertainty pixels indicate potential mislabeling or model disagreement, whereas low-uncertainty regions are more likely to be trustworthy.

4.1.4 Label Transformation Network (LTN)

4.1.4.1 Distilled Sample Selection via Pixel Entropy

Before estimating the instance-dependent transition matrix, we refine the noisy dataset by selecting a distilled subset $\mathcal{D}^* \subset \tilde{\mathcal{D}}$ with lower mislabel risk. Rather than thresholding only on global confidence, we compute **pixel-level entropy** in each model's output to detect local ambiguities:

- **Entropy Computation:** For each pixel i, define

$$H_i = -\sum_c p_{i,c} \ln(p_{i,c}),$$

  using class probabilities $\{p_{i,c}\}$ predicted by U-Net or nnU-Net.

- **Category Weighting:** Convert entropy values into class-wise weights.

- **Weighted Selection:** Merge the two models' (entropy-weighted) predictions. Pixels whose maximum probability surpasses a global average are retained in $\mathcal{D}^*$.

This process eliminates extremely uncertain or inconsistent pixels while preserving lesion-boundary pixels that show moderate confidence and consistent entropy patterns.

4.1.4.2 IDTM Parameterization and Training Objective

We denote by $T(x;\theta)$ the parametric mapping from "clean" to "noisy" labels. Let $\bar{y}^*$ be a refined (clean-like) prediction for a distilled sample x^*. Then the noisy label $\tilde{y}^*$ can be approximated by:

$$\tilde{y}^* = T^\top \bar{y}^*.$$

We adopt a *Rectified BCE* (RBCE) loss to handle partial label errors:

$$L_{\text{rbce}} = -\sum_{x^* \in \mathcal{D}^*} \sum_i \mathbf{1}[\tilde{y}_i^* = 1] \ln\big(\bar{y}_i^*\big). \tag{4.10}$$

To further reduce the impact of mislabeled or high-entropy pixels, an uncertainty re-weighting factor is used as follows

$$\psi_1(x) = \frac{\text{Var}_{\text{ensemble}}(x)}{\max_x \text{Var}_{\text{ensemble}}(x)} + \frac{H_{\text{pixel}}(x)}{\max_x H_{\text{pixel}}(x)} \tag{4.11}$$

$$L_{\text{ua}} = L_{\text{rbce}} \times \exp\big[-\psi_1(x^*)\big], \tag{4.12}$$

where ψ_1 measures local uncertainty by combining ensemble variance and pixel entropy. Consequently, the most dubious samples have exponentially lower influence during training.

Moreover, medical lesions are often sparse. To deter $T(x)$ from overfitting easy background pixels, we introduce a pixel entropy regularizer

$$L_{\text{entropy}} = -\sum_{c} p_c \, \ln(p_c), \tag{4.13}$$

where $\{p_c\}$ are the LTN's predicted class probabilities. Balancing these yields:

$$L_g \;=\; \sum_{x^* \in \mathcal{D}^*} w(x^*)\, L_{\text{rbce}}(\tilde{y}^*, T, \bar{y}^*) \;\; - \;\; \gamma\, L_{\text{entropy}}, \tag{4.14}$$

with $\gamma > 0$ modulating the strength of the entropy penalty, and $w(x^*)$ encapsulating uncertainty re-weighting.

4.1.5 Target Segmentation Network (TSN)

After obtaining $T(x;\theta)$, we train a final segmentation model $f(x;\phi)$ under a *forward correction* strategy. The TSN's "clean" prediction $\bar{y}$ is mapped to the noisy label space by:

$$\hat{y} \;=\; T^{\top}\, \bar{y}.$$

We then measure its discrepancy with the given noisy label $\tilde{y}$ via:

$$L_{\text{BCE}} \;=\; -\sum_{i} \Big[\, \tilde{y}_i \ln(\hat{y}_i) \;+\; (1-\tilde{y}_i) \ln\big(1-\hat{y}_i\big) \Big]. \tag{4.15}$$

Minimizing this loss with respect to ϕ trains the TSN to segment lesions robustly, harnessing the IDTM to mitigate label noise in the gradient signals.

TABLE 4.2: Details of the datasets used in the experiment.

Dataset Name	Source and Details	Image Size
Lab Infarct (private)	6096 images and masks	256×256
Brain Tumor[97]	3064 images and masks	256×256
Merging Brain Tumor[103]	2642 images and masks	256×256

4.2 Experiment

4.2.1 Datasets and Experimental Settings

4.2.1.1 Datasets

We utilized three datasets, two public datasets and one private dataset, the details of which are presented in Table 4.2. Ethical approval for the secondary use of the private dataset, which comprises 6,096 images, was obtained from the University of Technology Sydney (UTS) under approval number ETH24-10033.

The public dataset (Brain Tumor) [97] contains 3,064 images of brain tumors. Another public dataset (Merging Brain Tumor) [103], after we removed images without tumors and modified images and labels with tumors, includes 2,642 images of brain tumors. The training, validation, and test sets are split at a 7:1:2 ratio.

4.2.1.2 Noise Settings

To simulate expert annotation errors in 2D segmentation, we randomly perturb 15%–90% of training labels by applying either morphological operations (dilation or erosion with a 3–7 pixel structuring element) or affine transformations (random translation ± 5 pixels, rotation $\pm 10^{\circ}$, scaling 0.9–1.1). This introduces realistic, structured label noise mimicking common human annotation mistakes. Figure 2.1 shows a comparison between the original ground-truth masks (top row) and their corresponding noisy masks (bottom row), illustrating realistic annotation perturbations that mimic common human labeling errors.

4.2.1.3 Compared Methods

We use the nn-U-Net model architecture [104] as a baseline to evaluate three datasets with varying noise ratios. We then evaluate our proposed method on these datasets and compare its performance with state-of-the-art methods, including: the robust t-loss that employs a t-distribution-based loss to handle label noise [105]; a tri-network framework that achieves robust segmentation via mutual correction from non-expert labels [33]; a superpixel-guided iterative learning method that suppresses noise during training [98]; and a spatial correction approach that leverages spatial consistency to refine noisy annotations [106].

4.2.1.4 Evaluation Metrics

The Dice coefficient, also known as the Dice Similarity Coefficient (DSC), measures the similarity between two samples. It is especially useful for image segmentation, where it is commonly used to assess segmentation accuracy in medical imaging[99].

The Dice coefficient is calculated using the following formula:

$$\text{Dice} = \frac{2\,|A \cap B|}{|A| + |B|} \tag{4.16}$$

Intersection over Union (IoU) is a metric used to evaluate the overlap between predicted results and ground truth in object detection and image segmentation tasks. It measures how well the predicted region overlaps with the actual region [107].

The IoU is calculated using the following formula:

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|} \tag{4.17}$$

4.2.2 Implementation Details

The training setup comprised an Nvidia GeForce RTX 4090 graphics card, an Intel i9-13900KF CPU, and 64GB of DDR5 RAM running at 5200MHz. The training is exclusively on the Windows 11 operating system. In the proposed method, we use an Adam optimizer with a training batch size of 16. During Uncertainty Estimation Ensemble Module training, the initial learning rate is set to 1e-3, with a decay rate of 85% and a momentum value of 0.9. The training process consists of 100 epochs. For LTN, full training takes 100 epochs, and the learning rate is initialized to 1e-3, with a 1% degradation after every 3 epochs. For TSN, the full training takes 20 epochs, and the learning rate is initialized to 1e-3 decayed by 0.9 after each epoch. In the state-of-the-art methods, the full training takes 100 epochs, and the learning rate is initialized to 1e-3 with 1% degradation after every 3 epochs. Using Adam optimizer with a training batch size of 16.

4.3 Results

TABLE 4.3: Performance comparison of Lab Infarct dataset with different noise settings. The values are Dice and IoU scores.

	nn-Unet(baseline)[104]		T-loss[105]		Tri-net[108]		Superpixel[98]		Spatial Correction[106]		Ours	
Noise rate	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
15%	0.7997	0.6663	0.8087	0.6788	0.8300	0.7094	0.8338	0.7150	0.8425	0.7279	**0.8598**	**0.7541**
30%	0.7815	0.6414	0.7968	0.6623	0.8103	0.6811	0.8235	0.7000	0.8249	0.7020	**0.8447**	**0.7312**
45%	0.7548	0.6061	0.7876	0.6496	0.8085	0.6785	0.8152	0.6878	0.8211	0.6965	**0.8335**	**0.7145**
60%	0.7241	0.5675	0.7802	0.6397	0.8058	0.6748	0.8097	0.6803	0.8104	0.6813	**0.8225**	**0.6985**
75%	0.7013	0.5400	0.7773	0.6357	0.7829	0.6432	0.7966	0.6619	0.8005	0.6674	**0.8089**	**0.6791**
90%	0.6395	0.4701	0.7548	0.6062	0.7673	0.6225	0.7701	0.6262	0.7710	0.6273	**0.7971**	**0.6627**

4.3.1 Comparison of Lab Infarct Dataset

Table 4.3 compares the LUT framework with the baseline and state-of-the-art methods on the Lab Infarct dataset. The baseline uses nn-Unet without noise learning. LUT outperforms the baseline across all noise levels, improving Dice/IoU by 0.0601/0.0878 at 15%, 0.0632/0.0898 at 30%, 0.0787/0.1084 at 45%, 0.0984/0.131 at 60%, 0.1076/0.1391 at 75%, and 0.1576/0.1926 at 90%.

LUT also surpasses state-of-the-art methods, with Dice/IoU gains of 0.0423/0.0565 over T-loss, 0.0338/0.0402 over Tri-net, 0.027/0.0365 over Superpixel, and 0.0261/0.0354 over Spatial Correction at 90% noise. These results confirm LUT's superior segmentation accuracy and robustness.

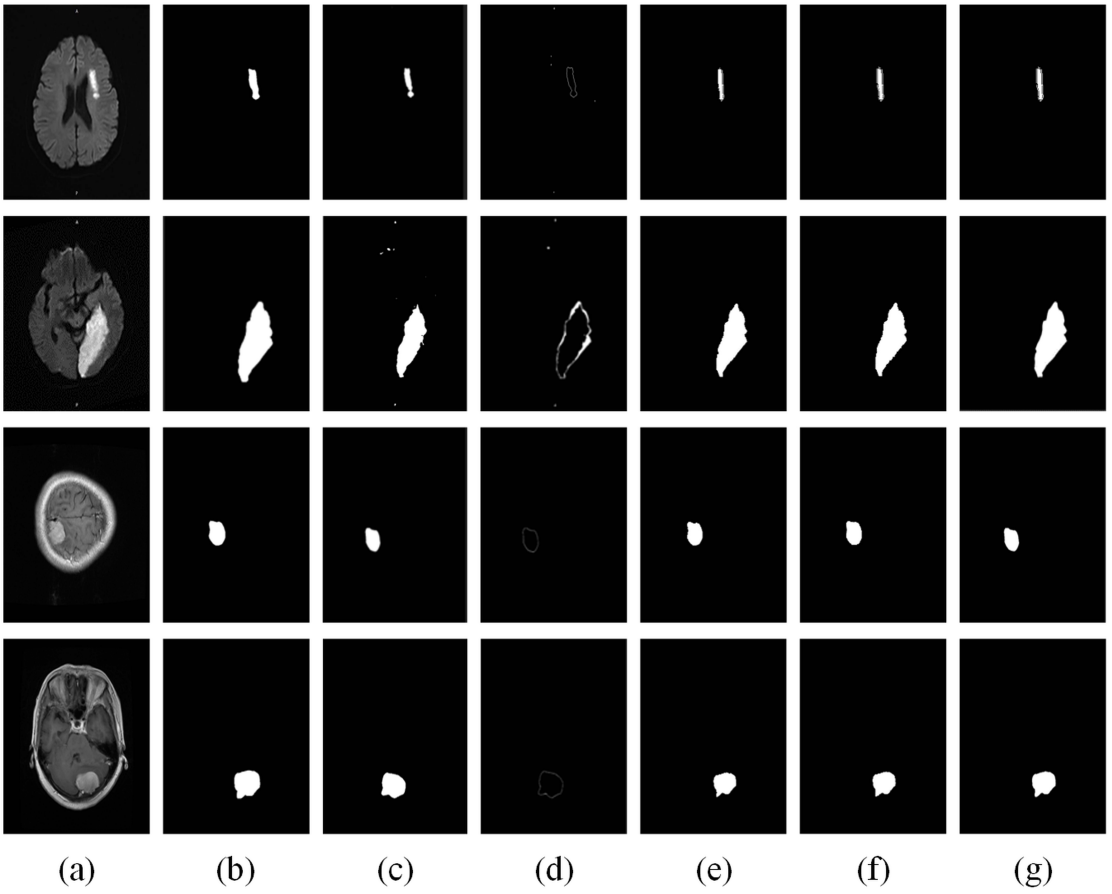


FIGURE 4.2: Prediction results of our method's components in the ablation experiments on the Lab Infarct and Brain Tumor dataset. (a) Training samples. (b) Noisy labels. (c) Initial results by UEEM. (d) Predictive uncertainty. (e) Original LTN. (f) +Pixel entropy regularizer. (g) +URW

TABLE 4.4: Performance comparison of Brain Tumor dataset with different noise settings. The values are Dice and IoU scores.

	nn-Unet(baseline)[104]		T-loss[105]		Tri-net[108]		Superpixel[98]		Spatial Correction[106]		Ours	
Noise setting	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
15%	0.8144	0.6869	0.8328	0.7135	0.8476	0.7355	0.8563	0.7338	0.8792	0.7844	**0.8917**	**0.8045**
30%	0.7991	0.6654	0.8195	0.6943	0.8241	0.7008	0.8340	0.7153	0.8577	0.7509	**0.8655**	**0.7629**
45%	0.7615	0.6148	0.8080	0.6778	0.8156	0.6887	0.8264	0.7042	0.8449	0.7315	**0.8563**	**0.7488**
60%	0.7391	0.5862	0.7885	0.6509	0.7952	0.6600	0.8192	0.6937	0.8396	0.7235	**0.8469**	**0.7345**
75%	0.7208	0.5634	0.7823	0.6424	0.7908	0.6540	0.8065	0.6757	0.8208	0.6960	**0.8340**	**0.7153**
90%	0.6638	0.4968	0.7621	0.6156	0.7694	0.6253	0.7813	0.6411	0.8006	0.6675	**0.8145**	**0.6871**

4.3.2 Comparison with Brain Tumor Dataset

Table 4.4 compares the LUT framework with the baseline and state-of-the-art methods on the Brain Tumor dataset. LUT consistently outperforms the baseline, which lacks noise learning, across all noise levels. Specifically, Dice/IoU improvements are 0.0773/0.1176 at 15%, 0.0664/0.0975 at 30%, 0.0948/0.134 at 45%, 0.1078/0.1483 at 60%, 0.1132/0.1519 at 75%, and 0.1507/0.1903 at 90%.

LUT also surpasses state-of-the-art methods at all noise levels. At 90% noise, Dice/IoU improvements are 0.0624/0.0715 over T-loss, 0.0451/0.0618 over Tri-net, 0.0332/0.046 over Superpixel, and 0.0139/0.0196 over Spatial Correction. These results confirm LUT's superior segmentation accuracy and robustness.

TABLE 4.5: Performance comparison of Merging Brain Tumor dataset with different noise settings. The values are Dice and IoU scores.

	nn-Unet(baseline)[104]		T-loss[105]		Tri-net[108]		Superpixel[98]		Spatial Correction[106]		Ours	
Noise setting	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
15%	0.8198	0.6947	0.8404	0.7248	0.8559	0.7481	0.8635	0.7598	0.8848	0.7933	**0.9011**	**0.8199**
30%	0.8027	0.6704	0.8293	0.7083	0.8350	0.7167	0.8458	0.7328	0.8646	0.7615	**0.8820**	**0.7899**
45%	0.7724	0.6292	0.8116	0.6830	0.8253	0.7025	0.8352	0.7170	0.8559	0.7481	**0.8675**	**0.7660**
60%	0.7409	0.5885	0.7910	0.6543	0.8027	0.6704	0.8212	0.6966	0.8423	0.7275	**0.8530**	**0.7436**
75%	0.7275	0.5717	0.7866	0.6482	0.7905	0.6605	0.8124	0.6840	0.8235	0.7000	**0.8402**	**0.7245**
90%	0.6674	0.5008	0.7665	0.6214	0.7737	0.6310	0.7899	0.6528	0.8041	0.6723	**0.8182**	**0.6924**

4.3.3 Comparison of Merging Brain Tumor Dataset

Table 4.5 compares the LUT framework with the baseline and state-of-the-art methods on the Merging Brain Tumor dataset. LUT consistently outperforms the baseline across all noise levels. At 90% noisy labels, the Dice score improves by 0.1508 and the IoU by 0.1916, demonstrating superior prediction accuracy. LUT also surpasses state-of-the-art methods at all noise levels. At 90% noise, Dice/IoU improvements are 0.0517/0.071 over T-loss, 0.0445/0.0614 over Tri-net, and 0.0283/0.0396 over Superpixel. These results confirm LUT's robustness and superior segmentation performance.

4.3.4 Ablation Experiment

In this section, we evaluate the individual contributions of each component of the proposed methodology through ablation experiments. We categorize the noise ratios into three classes: small (15%), moderate (45%), and severe (75%). These representative noise levels are used to assess the model's performance and the robustness of each module under different noise conditions.

TABLE 4.6: Ablation experiment on the Lab Infarct dataset with 15%, 45%, and 75% noisy data.

	15% noisy data		**45% noisy data**		**75% noisy data**	
	Dice↑	**IoU↑**	**Dice↑**	**IoU↑**	**Dice↑**	**IoU↑**
Baseline	0.7997	0.6663	0.7548	0.6061	0.7013	0.5400
LTN Strategy						
+ LTN	0.8367	0.7192	0.7750	0.6326	0.7371	0.5836
+ PE regularizer	0.8453	0.7321	0.8053	0.6740	0.7908	0.6540
+ URW	**0.8598**	**0.7541**	**0.8335**	**0.7145**	**0.8089**	**0.6791**

TABLE 4.7: Ablation experiment on the Brain Tumor dataset with 15%, 45%, and 75% noisy data.

	15% noisy data		**45% noisy data**		**75% noisy data**	
	Dice↑	**IoU↑**	**Dice↑**	**IoU↑**	**Dice↑**	**IoU↑**
Baseline	0.8144	0.6869	0.7615	0.6148	0.7208	0.5634
LTN Strategy						
+ LTN	0.8634	0.7596	0.7842	0.6450	0.7640	0.6181
+ PE regularizer	0.8826	0.7899	0.8360	0.7182	0.8109	0.6819
+ URW	**0.8917**	**0.8045**	**0.8563**	**0.7488**	**0.8340**	**0.7153**

TABLE 4.8: Ablation experiment on the Merging Brain Tumor dataset with 15%, 45%, and 75% noisy data.

	15% noisy data		**45% noisy data**		**75% noisy data**	
	Dice↑	**IoU↑**	**Dice↑**	**IoU↑**	**Dice↑**	**IoU↑**
Baseline	0.8198	0.6947	0.7724	0.6292	0.7275	0.5717
LTN Strategy						
+ LTN	0.8684	0.7675	0.7897	0.6526	0.7665	0.6214
+ PE regularizer	0.8881	0.7987	0.8423	0.7275	0.8151	0.6879
+ URW	**0.9011**	**0.8199**	**0.8675**	**0.7660**	**0.8402**	**0.7245**

4.3.4.1 Impact of original LTN without UA Re-Weighting and pixel entropy regularizer

To assess the original LTN, we compared it with the baseline across three datasets. For Lab Infarct (Table 4.6), Dice/IoU increased by 0.037/0.0529 at 15%, 0.0202/0.0265 at 45%, and 0.0358/0.0436 at 75%. For Brain Tumor (Table 4.7), Dice/IoU improved by 0.049/0.0727 at 15%, 0.0227/0.0302 at 45%, and 0.0432/0.0547 at 75%. For Merging Brain Tumor (Table 4.8), Dice/IoU increased by 0.0486/0.0728 at 15%, 0.0173/0.0234 at 45%, and 0.039/0.0497 at 75%. These results highlight the effectiveness of the original LTN in improving segmentation performance across datasets and noise levels.

4.3.4.2 The impact of pixel entropy regularizer(PE regularizer)

To assess the impact of the pixel entropy regularizer, we integrated it into LTN training, yielding consistent improvements across datasets. For the Lab Infarct dataset (Table 4.6), Dice and IoU increased by 0.0086/0.0129 at 15%, 0.0303/0.0414 at 45%, and 0.0537/0.0704 at 75%. For the Brain Tumor dataset (Table 4.7), Dice and IoU improved by 0.0192/0.0303 at 15%, 0.0518/0.0732 at 45%, and 0.0469/0.0638 at 75%. For the Merging Brain Tumor dataset (Table 4.8), Dice and IoU increased by 0.0197/0.0312 at 15%, 0.0524/0.0747 at 45%, and 0.0486/0.0665 at 75%. These results validate the effectiveness of the pixel entropy regularizer in enhancing LTN performance across varying noise levels.

4.3.4.3 The impact of Uncertainty Re-weighting (URW)

To assess the impact of Uncertainty Re-weighting (URW), we integrated it into LTN training. For the Lab Infarct dataset (Table 4.6), URW improved performance over the PE-regularized model. At 15% noise, the Dice score increased by 0.0145 and IoU by 0.022; at 45%, by 0.0282 and 0.0405; at 75%, by 0.0181 and 0.0251. For the Brain Tumor dataset (Table 4.7), URW also enhanced performance. At 15% noise, Dice and IoU increased by 0.0091 and 0.0146; at 45%, by 0.0203 and 0.0306; at 75%, by 0.0231 and 0.0334. For the Merging Brain Tumor dataset (Table 4.8), URW showed similar benefits. At 15% noise, Dice and IoU improved by 0.013 and 0.0212; at 45%, by 0.0252 and 0.0385; at 75%, by

0.0251 and 0.0366. These results confirm URW's effectiveness in further enhancing LTN performance. The experimental results across three datasets confirm the effectiveness of each proposed component in improving segmentation performance under noisy labels. The original LTN consistently outperforms the baseline, demonstrating its robustness. The integration of the pixel entropy regularizer further enhances performance, particularly at higher noise levels. Additionally, Uncertainty Re-weighting (URW) further improves the PE regularizer, reinforcing its impact. Figure 4.2 visualizes the model's performance improvements through multiple stages, demonstrating the effectiveness of the proposed method.

4.3.5 Pixel Entropy Regularizer Parameter Tuning

We emphasize that all hyperparameter tuning, including the selection of the regularization parameter γ, was conducted exclusively on the validation sets to prevent information leakage and ensure an unbiased evaluation of model performance. Considering the intrinsic differences among datasets, distinct γ value ranges were explored, tailored to each dataset's characteristics. For each dataset, five carefully selected γ values were evaluated, sufficiently covering the performance peak, thereby demonstrating that the optimal γ parameter was captured within the tested range.

The parameter γ controls the strength of the constraint on the trajectory of the instance-dependent transition matrix (IDTM). We evaluate its effect under both *with* and *without* uncertainty-aware (UA) re-weighting conditions. Quantitative results are summarized in Figures 4.3 to 4.8.

Without UA re-weighting, γ increases with the noise rate and varies across datasets. At 15% noise, γ is set to 0.02 for all three datasets. At 30% noise, γ is 0.03 (Lab Infarct), 0.04 (Brain Tumor), and 0.05 (Merging Brain Tumor). At 45%, it rises to 0.08, 0.07, and 0.06, respectively. For 60% noise, γ values are 0.10, 0.09, and 0.08. At 75%, they increase to 0.19, 0.18, and 0.16. Finally, at 90% noise, γ is 0.26 (Lab Infarct), 0.23 (Brain Tumor), and 0.22 (Merging Brain Tumor).

Under UA re-weighting, a similar trend is observed with generally larger γ values. At 15% noise, all datasets use $\gamma = 0.04$. At 30%, values are 0.06 (Lab Infarct), 0.05 (Brain Tumor), and 0.04 (Merging Brain Tumor). For 45%, they increase to 0.07, 0.06, and 0.05, respectively. At 60%, γ values are 0.09 (Lab Infarct), 0.08 (Brain Tumor), and 0.07 (Merging Brain Tumor). For 75%, the values are 0.17, 0.15, and 0.13, and at 90%, they reach 0.24 (Lab Infarct), 0.21 (Brain Tumor), and 0.19 (Merging Brain Tumor).

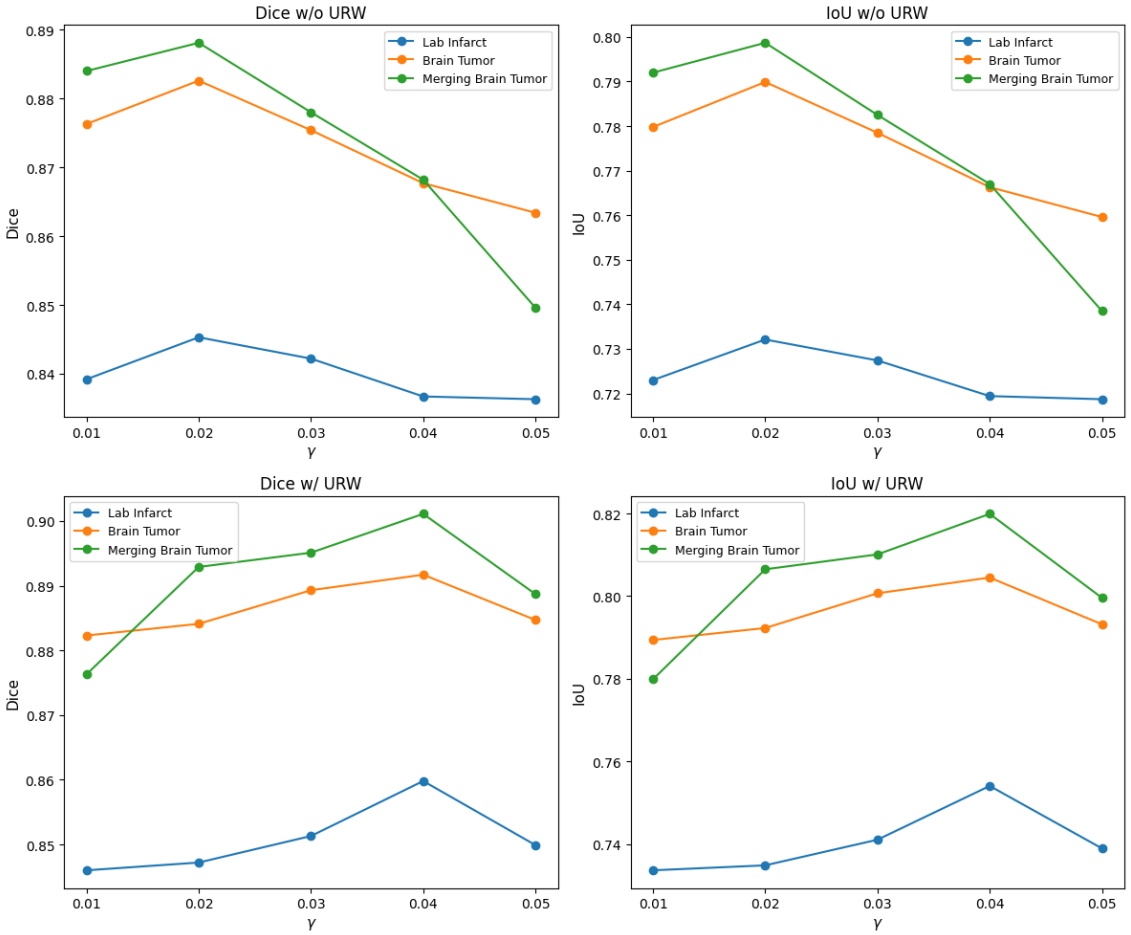


FIGURE 4.3: Dice and IoU performance across different values of the regularization parameter γ on the three datasets under 15% label noise.

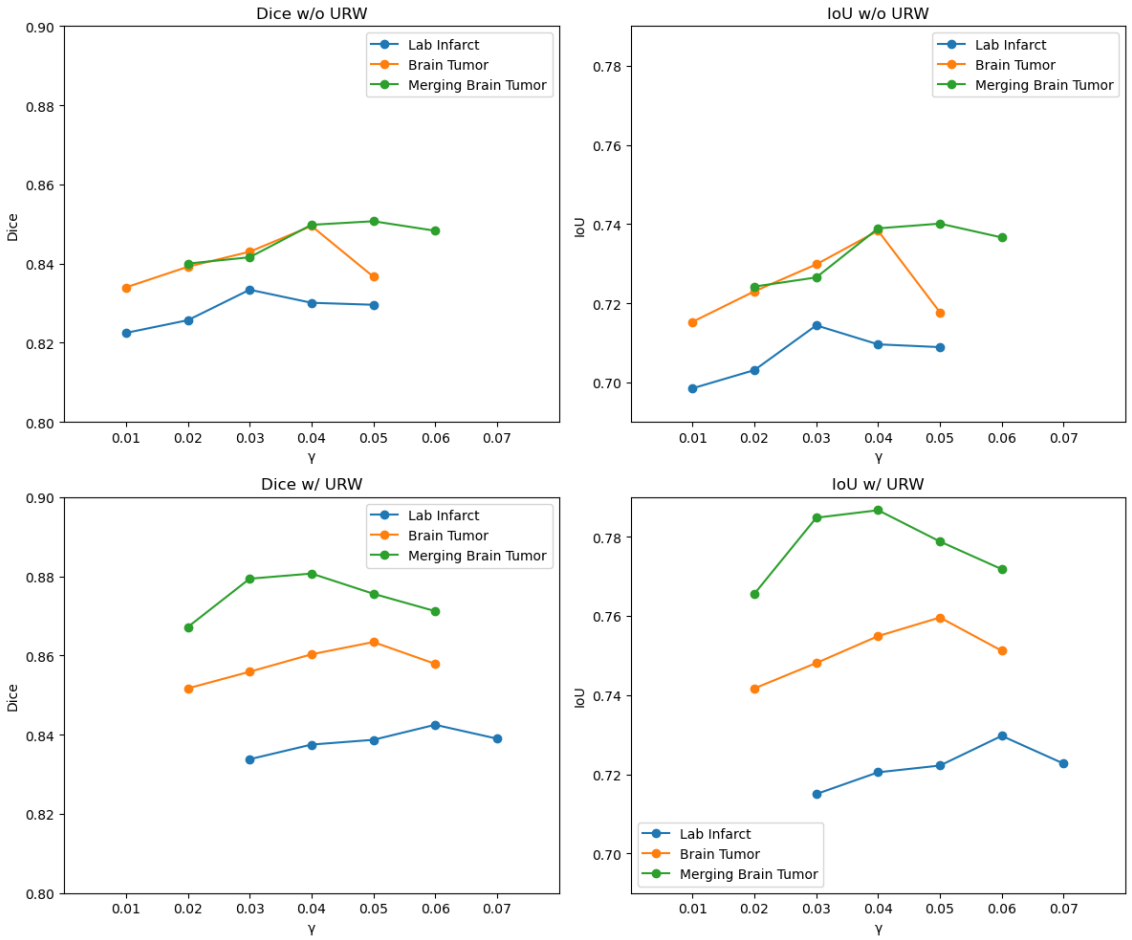


FIGURE 4.4: Dice and IoU performance across different values of the regularization parameter γ on the three datasets under 30% label noise.

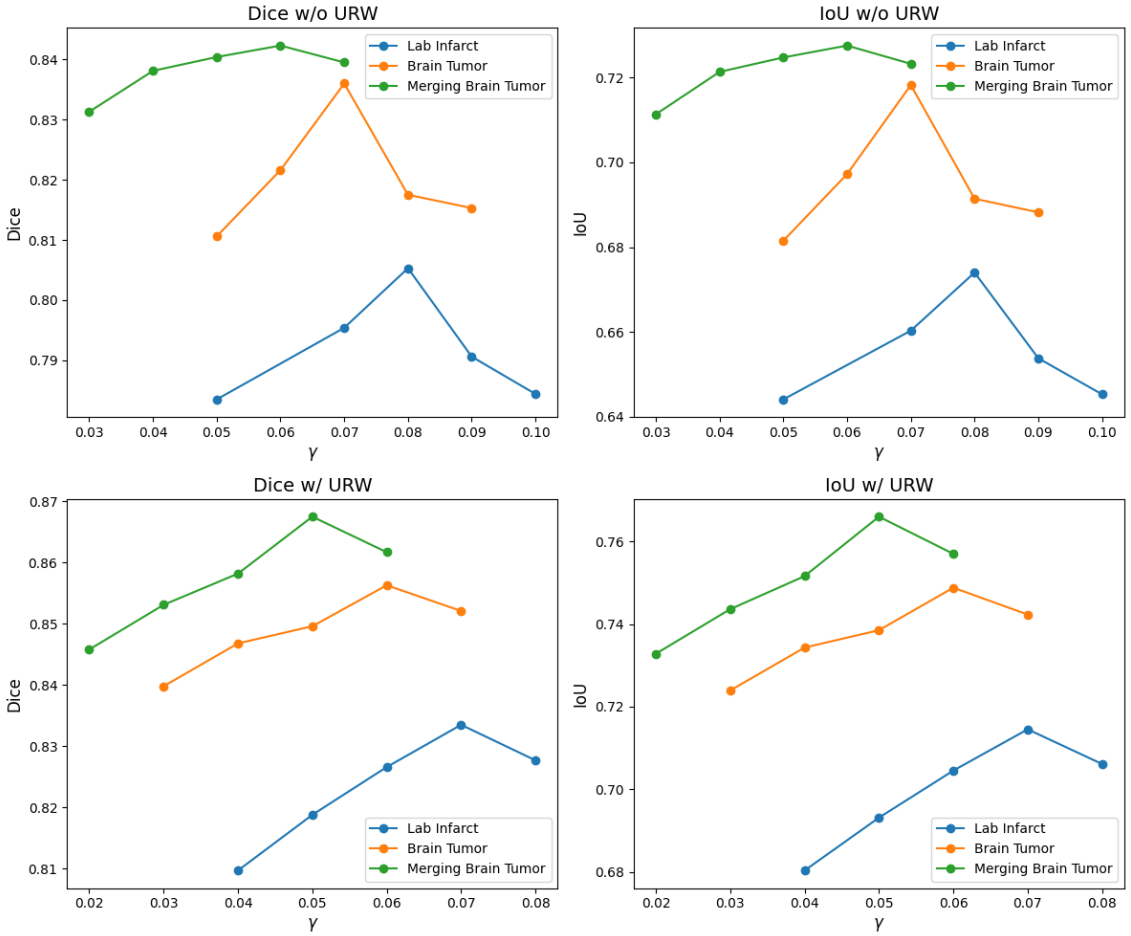


FIGURE 4.5: Dice and IoU performance across different values of the regularization parameter γ on the three datasets under 45% label noise.

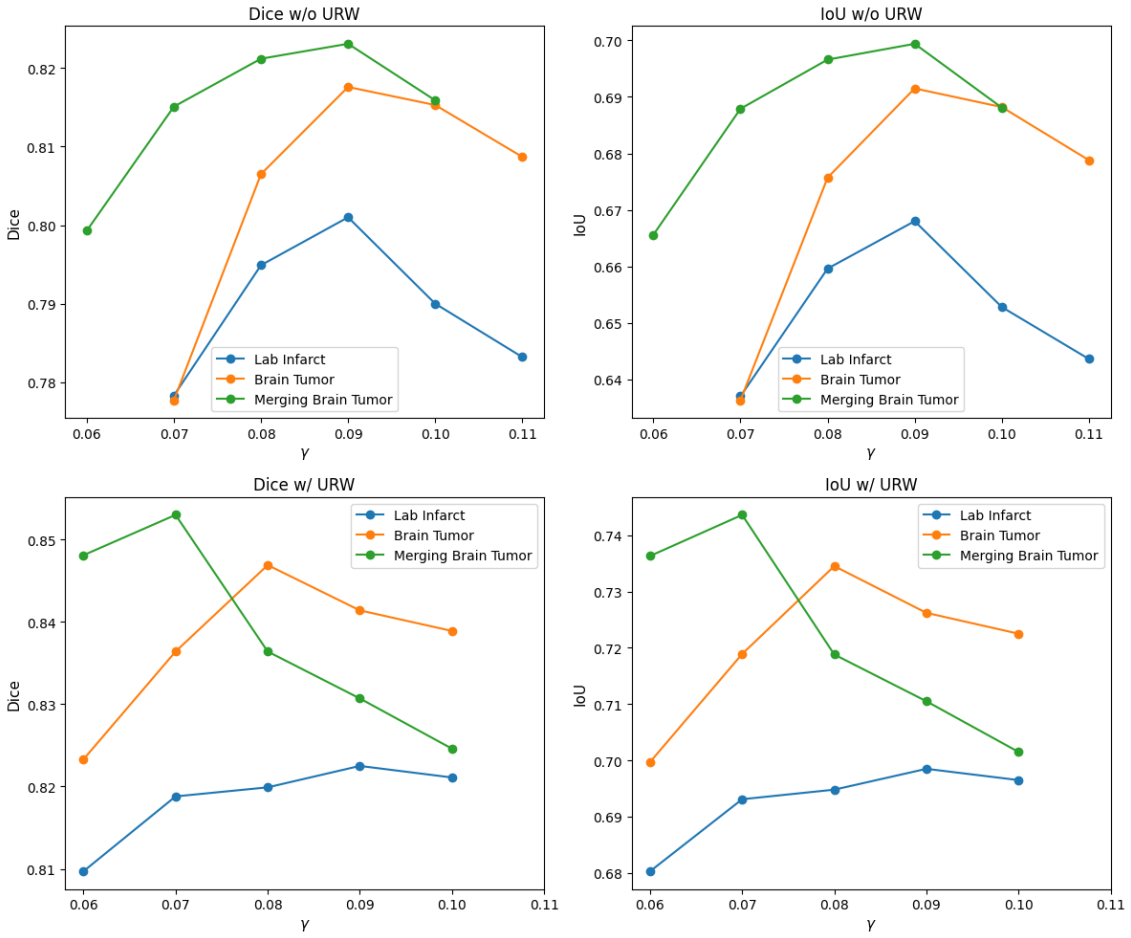


FIGURE 4.6: Dice and IoU performance across different values of the regularization parameter γ on the three datasets under 60% label noise.

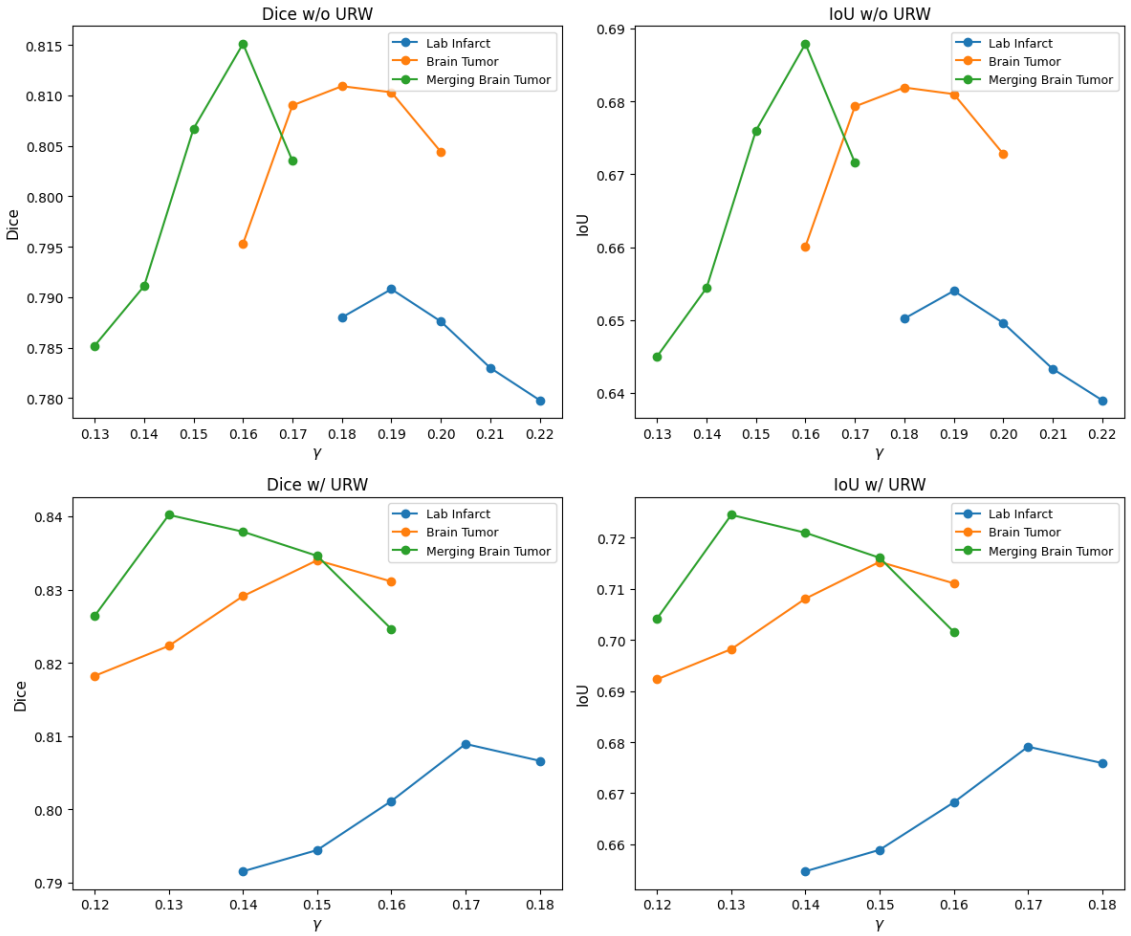


FIGURE 4.7: Dice and IoU performance across different values of the regularization parameter γ on the three datasets under 75% label noise.

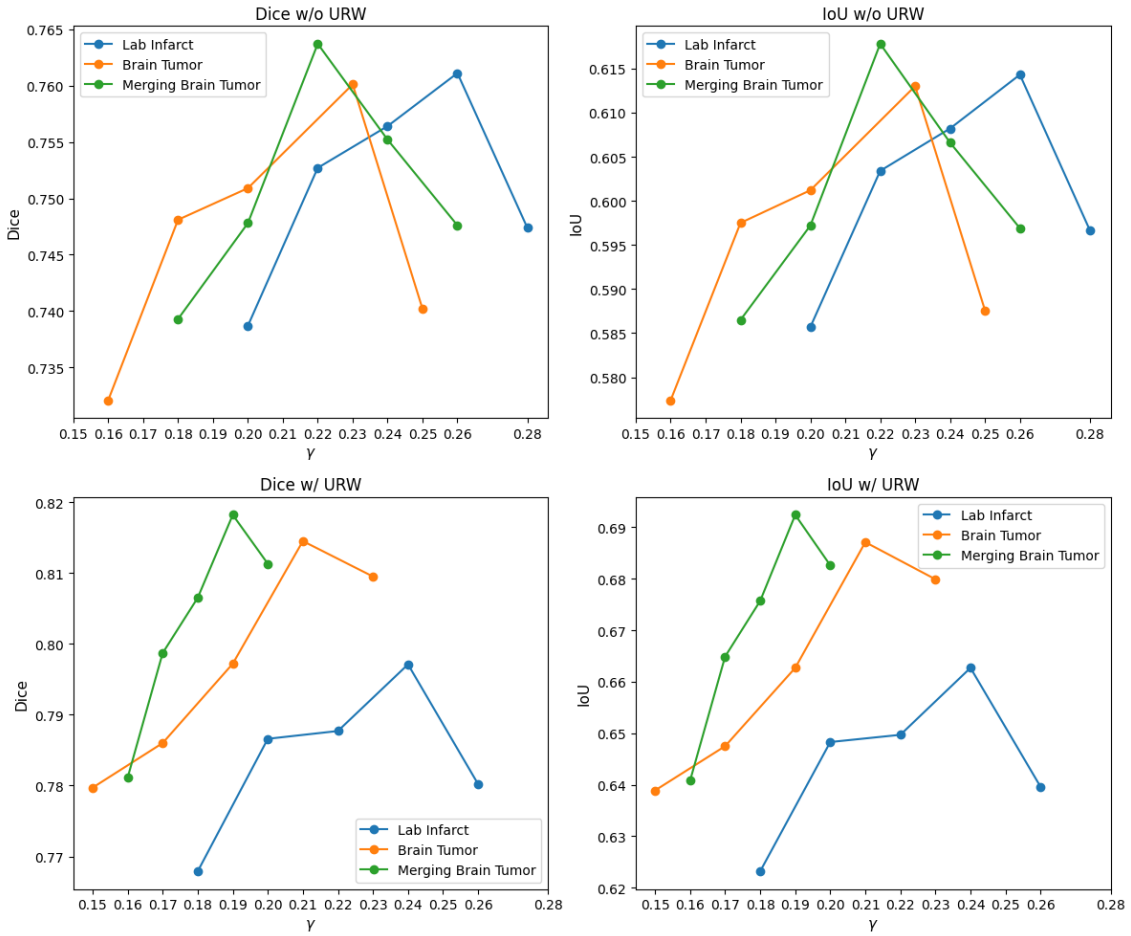


FIGURE 4.8: Dice and IoU performance across different values of the regularization parameter γ on the three datasets under 90% label noise.

Chapter 5

Conclusion and Future Work

This thesis addresses two fundamental challenges in medical image analysis: class imbalance in disease classification and label noise in image segmentation.

To tackle the first challenge, we conduct a comprehensive evaluation of self-supervised pretrained models, particularly DINOv2, across multiple class-imbalanced medical imaging datasets. The results consistently show that DINOv2 outperforms conventional supervised models such as ResNet, VGG, and ViT, demonstrating enhanced generalization and robustness. Building on this, we propose the Self-Calibrated Feature Denoising (SCFD) framework, which selectively suppresses noisy representations from minority classes while preserving the pretrained backbone's structural integrity. This design significantly enhances classification accuracy and feature discriminability, underscoring its potential for deployment in clinical settings.

To address label noise in segmentation, we introduce the Label Uncertainty Transformation (LUT) framework, a statistically grounded method composed of three core components that jointly mitigate the detrimental effects of noisy supervision. Extensive experiments on three noisy brain imaging datasets validate the framework's effectiveness and robustness. Ablation studies further confirm the individual contributions of each module.

Future work will explore several directions to enhance the proposed methods. For classification, we aim to develop more adaptive denoising strategies and to incorporate domain-specific self-supervised objectives to better capture the characteristics of minority classes.

Leveraging multi-modal imaging and clinical priors may further improve robustness and interpretability. In segmentation, future efforts will focus on refining uncertainty estimation, exploring alternative network architectures, and improving the modeling of label transitions. We also plan to use unlabeled data through self-supervised or transfer learning techniques to reduce reliance on noisy annotations.

Finally, all methods will be evaluated on larger and more diverse clinical datasets to assess generalization across different imaging protocols and institutions. These directions aim to further advance the robustness, scalability, and clinical utility of automated medical image analysis systems.

Bibliography

[1] Hamidreza Saber, Amanda G Thrift, Moira K Kapral, Ashkan Shoamanesh, Amin Amiri, Mohammad T Farzadfard, Réza Behrouz, and Mahmoud Reza Azarpazhooh. "Incidence, recurrence, and long-term survival of ischemic stroke subtypes: a population-based study in the Middle East". In: *International Journal of Stroke* 12.8 (2017), pp. 835–843. DOI: 10.1177/1747493016684843.

[2] Wei Wu, Yi Guan, Kan Xu, Xi-Jia Fu, Xiao-Feng Lei, Li-Jian Lei, Zhi-Qing Zhang, Yan Cheng, and Yun-Qian Li. "Plasma Homocysteine Levels Predict the Risk of Acute Cerebral Infarction in Patients with Carotid Artery Lesions". In: *Molecular Neurobiology* 53 (2016), pp. 2510–2517. DOI: 10.1007/s12035-015-9226-y.

[3] Huikai Shao, Xia He, Lijuan Zhang, Shan Du, Xiaoqing Yi, Xiaojiao Cui, Xinxia Liu, Shengfeng Huang, and Rongsheng Tong. "Efficacy of ligustrazine injection as adjunctive therapy in treating acute cerebral infarction: a systematic review and meta-analysis". In: *Frontiers in Pharmacology* 12 (2021), p. 761722. DOI: 10.3389/fphar.2021.761722.

[4] Cancer.Net. *Brain Tumor: Statistics*. https://www.cancer.net/cancer-types/brain-tumor/statistics. Accessed March 2023. 2023.

[5] Mayo Clinic. *Brain tumor - Symptoms and causes*. https://www.mayoclinic.org/diseases-conditions/brain-tumor/symptoms-causes/syc-20350084. Accessed April 21, 2023. 2023.

[6] American Cancer Society. *Lung cancer. Information and resources about cancer: Breast, colon, lung, prostate, skin*. https://www.cancer.org/cancer/types/lung-cancer.html. 2023.

[7] Bhawna Goyal, Sunil Agrawal, and BS Sohi. "Noise issues prevailing in various types of medical images". In: *Biomedical & Pharmacology Journal* 11.3 (2018), p. 1227. DOI: 10.13005/bpj/1484.

[8] Weibin Wang, Dong Liang, Qingqing Chen, Yutaro Iwamoto, Xian-Hua Han, Qiaowei Zhang, Hongjie Hu, Lanfen Lin, and Yen-Wei Chen. "Medical Image Classification Using Deep Learning". In: *Deep Learning in Healthcare: Paradigms and Applications*. Ed. by Yen-Wei Chen and Lakhmi C. Jain. Cham: Springer International Publishing, 2020, pp. 33–51. DOI: 10.1007/978-3-030-32606-7_3.

[9] Saravanan Srinivasan, Prabin Selvestar Mercy Bai, Sandeep Kumar Mathivanan, Venkatesan Muthukumaran, Jyothi Chinna Babu, and Lucia Vilcekova. "Grade classification of tumors from brain magnetic resonance images using a deep learning technique". In: *Diagnostics* 13.6 (2023), p. 1153. DOI: 10.3390/diagnostics13061153.

[10] Didem Cifci, Gregory P Veldhuizen, Sebastian Foersch, and Jakob Nikolas Kather. "AI in computational pathology of cancer: improving diagnostic workflows and clinical outcomes?" In: *Annual Review of Cancer Biology* 7.1 (2023), pp. 57–71. DOI: 10.1146/annurev-cancerbio-061521-092038.

[11] Yaw George Boafo. "An overview of computer—aided medical image classification". In: *Multimedia Tools and Applications* (2024), pp. 1–39. DOI: 10.1007/s11042-024-19558-1.

[12] Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K. Nandi. "Medical image segmentation using deep learning: A survey". In: *IET Image Processing* 16.5 (2022), pp. 1243–1267. DOI: 10.1049/ipr2.12419.

[13] Majid Harouni, Mohsen Karimi, and Shadi Rafieipour. "Precise segmentation techniques in various medical images". In: *Artificial Intelligence and Internet of Things* (2021), pp. 117–166.

[14] Shah Foysal Hossain, Md Al Amin, Irin Akter Liza, Shahriar Ahmed, Md Musa Haque, Md Azharul Islam, and Sarmin Akter. "AI-Based Brain MRI Segmentation for Early Diagnosis and Treatment Planning of Low-Grade Gliomas in the USA". In: *British Journal of Nursing Studies* 3.2 (2023), pp. 37–55. DOI: 10.32996/bjns.2023.3.2.5.

[15] Mathieu Hatt, John A Lee, Charles R Schmidtlein, Issam El Naqa, Curtis Caldwell, Elisabetta De Bernardi, Wei Lu, Shiva Das, Xavier Geets, Vincent Gregoire, et al. "Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group No. 211". In: *Medical physics* 44.6 (2017), e1–e42.

[16] K. K. D. Ramesh, G. K. Kumar, K. Swapna, D. Datta, and S. S. Rajest. "A Review of Medical Image Segmentation Algorithms". In: *EAI Endorsed Transactions on Pervasive Health and Technology* 7.27 (2021), e2. DOI: 10.4108/eai.13-7-2018.165523.

[17] Jiaxin Zhuang, Jiabin Cai, Ruixuan Wang, Jianguo Zhang, and Weishi Zheng. "Care: Class attention to regions of lesion for classification on imbalanced data". In: *International Conference on Medical Imaging with Deep Learning*. PMLR. 2019, pp. 588–597.

[18] Mohammed Temraz and Mark T. Keane. "Solving the class imbalance problem using a counterfactual method for data augmentation". In: *Machine Learning with Applications* 9 (2022), p. 100375. DOI: 10.1016/j.mlwa.2022.100375.

[19] Salim Rezvani and Xizhao Wang. "A broad review on class imbalance learning techniques". In: *Applied Soft Computing* 143 (2023), p. 110415.

[20] Anil K. Jain, Debayan Deb, and J. J. Engelsma. "Biometrics: Trust, but Verify". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 4.3 (2021), pp. 303–323. DOI: 10.1109/TBIOM.2021.3096727.

[21] M. Abdelhamid and A. Desai. "Balancing the Scales: A Comprehensive Study on Tackling Class Imbalance in Binary Classification". In: *arXiv preprint arXiv:2409.19751* (2024). arXiv: 2409.19751 [cs.LG].

[22] R. Suguna, J. Suriya Prakash, H. Aditya Pai, et al. "Mitigating class imbalance in churn prediction with ensemble methods and SMOTE". In: *Scientific Reports* 15.1 (2025), p. 16256. DOI: 10.1038/s41598-025-01031-0.

[23] Xiaoling Gao, Muhammad Izzad Ramli, Marshima Mohd Rosli, Nursuriati Jamil, and Syed Mohd Zahid Syed Zainal Ariffin. "Revisiting self-supervised contrastive learning for imbalanced classification." In: *International Journal of Electrical &*

Computer Engineering (2088-8708) 15.2 (2025). DOI: 10.11591/ijece.v15i2.pp1949-1960.

[24] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. "Pre-trained models: Past, present and future". In: *Ai Open* 2 (2021), pp. 225–250. DOI: 10.1016/j.aiopen.2021.08.002.

[25] Biswajit Jena, Gopal Krishna Nayak, and Sanjay Saxena. "Convolutional neural network and its pretrained models for image classification and object detection: A survey". In: *Concurrency and Computation: Practice and Experience* 34.6 (2022), e6767.

[26] Nikolaos Giakoumoglou, Tania Stathaki, and Athanasios Gkelias. "A Review on Discriminative Self-supervised Learning Methods in Computer Vision". In: *arXiv preprint arXiv:2405.04969* (2024).

[27] Md. Eshmam Rayed, S.M. Sajibul Islam, Sadia Islam Niha, Jamin Rahman Jim, Md Mohsin Kabir, and M.F. Mridha. "Deep learning for medical image segmentation: State-of-the-art advancements and challenges". In: *Informatics in Medicine Unlocked* 47 (2024), p. 101504. ISSN: 2352-9148. DOI: 10.1016/j.imu.2024.101504.

[28] S. Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S. Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L. Prince, Daniel Rueckert, and Ronald M. Summers. "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises". In: *Proceedings of the IEEE* 109.5 (2021), pp. 820–838. DOI: 10.1109/JPROC.2021.3054390.

[29] Tim Rädsch, Annika Reinke, Vivienn Weru, Minu D Tizabi, Nicholas Schreck, A Emre Kavur, Bünyamin Pekdemir, Tobias Roß, Annette Kopp-Schneider, and Lena Maier-Hein. "Labelling instructions matter in biomedical image analysis". In: *Nature Machine Intelligence* 5.3 (2023), pp. 273–283. DOI: 10.1038/s42256-023-00625-5.

[30] Tatenda Magadza and Samuel Viriri. "Deep learning for brain tumor segmentation: A survey of state-of-the-art". In: *Journal of Imaging* 7.2 (2021), p. 19.

[31] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. "Learning from noisy labels with deep neural networks: A survey". In: *IEEE Transactions on Neural Networks and Learning Systems* 34.11 (2022), pp. 8135–8153. DOI: 10.1109/TNNLS.2022.3152527.

[32] Davood Karimi, Haoran Dou, Simon K. Warfield, and Ali Gholipour. "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis". In: *Medical Image Analysis* 65 (2020), p. 101759. DOI: 10.1016/j.media.2020.101759.

[33] Justin M. Johnson and Taghi M. Khoshgoftaar. "Survey on deep learning with class imbalance". In: *Journal of Big Data* 6.1 (2019), pp. 1–54.

[34] Lixin Gao, Lijun Zhang, Chun Liu, and Shaohui Wu. "Handling imbalanced medical image data: A deep-learning-based one-class classification approach". In: *Artificial intelligence in medicine* 108 (2020), p. 101935.

[35] Na Liu, Xiaomei Li, Ershi Qi, Man Xu, Ling Li, and Bo Gao. "A novel ensemble learning paradigm for medical diagnosis with imbalanced data". In: *IEEE Access* 8 (2020), pp. 171263–171280. DOI: 10.1109/ACCESS.2020.3014362.

[36] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura. "Handling Imbalanced Medical Datasets: Review of a Decade of Research". In: *Artificial Intelligence Review* 57.10 (2024), p. 273. DOI: 10.1007/s10462-023-10453-x.

[37] Annelies Carriero, Kim Luijken, Andries de Hond, Karel G. Moons, Ben van Calster, and Maarten van Smeden. "The harms of class imbalance corrections for machine learning based prediction models: a simulation study". In: *Statistics in Medicine* 44.3-4 (2025), e10320.

[38] Dong Peng, Tao Gu, Xiaolin Hu, and Chun Liu. "Addressing the multi-label imbalance for neural networks: An approach based on stratified mini-batches". In: *Neurocomputing* 435 (2021), pp. 91–102.

[39] Amadi G. Udu, Marwah T. Salman, Maryam K. Ghalati, Andrea Lecchini-Visintini, David R. Siddle, and Hongbiao Dong. "Emerging SMOTE and GAN-variants for Data Augmentation in Imbalance Machine Learning Tasks: A Review". In: *IEEE Access* (2025). DOI: 10.1109/ACCESS.2025.3584532.

[40] Liyuan Zhang, Romer Tanno, Mengchao Charles Xu, Cheng Jin, Jan Jacob, Oliver Cicarrelli, and Daniel Alexander. "Disentangling Human Error from Ground Truth in Segmentation of Medical Images". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. 2020, pp. 15750–15762.

[41] Ravi Ranjan Kumar and Rahul Priyadarshi. "Denoising and Segmentation in Medical Image Analysis: A Comprehensive Review on Machine Learning and Deep Learning Approaches". In: *Multimedia Tools and Applications* (2024), pp. 1–59. DOI: 10.1007/s11042-024-19313-6.

[42] Jialin Shi, Kailai Zhang, Chenyi Guo, Youquan Yang, Yali Xu, and Ji Wu. "A survey of label-noise deep learning for medical image analysis". In: *Medical image analysis* 95 (2024), p. 103166. DOI: 10.1016/j.media.2024.103166.

[43] Guo Zheng, Ahmed H. Awadallah, and Susan Dumais. "Meta Label Correction for Noisy Label Learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 12. 2021, pp. 11053–11061.

[44] Ali Hakami. "Strategies for overcoming data scarcity, imbalance, and feature selection challenges in machine learning models for predictive maintenance". In: *Scientific Reports* 14.1 (2024), p. 9645.

[45] Tiago Cortinhal, George Tzelepis, and Erdal Erdal Aksoy. "SalsaNext: Fast, Uncertainty-Aware Semantic Segmentation of LiDAR Point Clouds". In: *International Symposium on Visual Computing*. Cham: Springer International Publishing, 2020, pp. 207–222.

[46] Zhenzhen Wang, Carla Saoud, Sintawat Wangsiricharoen, Aaron W. James, Aleksander S. Popel, and Jeremias Sulam. "Label cleaning multiple instance learning: Refining coarse annotations on single whole-slide images". In: *IEEE Transactions on Medical Imaging* 41.12 (Dec. 2022), pp. 3952–3968. DOI: 10.1109/TMI.2022.3202759.

[47] Mayuri S. Shelke, Prashant R. Deshmukh, and Vijaya K. Shandilya. "A review on imbalanced data handling using undersampling and oversampling technique". In: *Int. J. Recent Trends Eng. Res* 3.4 (2017), pp. 444–449.

[48] Imane Araf, Ali Idri, and Ikram Chairi. "Cost-sensitive learning for imbalanced medical data: a review". In: *Artificial Intelligence Review* 57.4 (2024), p. 80. DOI: 10.1007/s10462-023-10652-8.

[49] Juan Terven, Diana-Margarita Cordova-Esparza, Julio-Alejandro Romero-González, Alfonso Ramírez-Pedraza, and EA Chávez-Urbiola. "A comprehensive survey of loss functions and metrics in deep learning". In: *Artificial Intelligence Review* 58.7 (2025), p. 195. DOI: 10.1007/s10462-025-11198-7.

[50] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. "A survey on self-supervised learning: Algorithms, applications, and future trends". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.12 (2024), pp. 9052–9071. DOI: 10.1109/TPAMI.2024.3415112.

[51] Jin Gao, Shubo Lin, Shaoru Wang, Yutong Kou, Zeming Li, Liang Li, Congxuan Zhang, Xiaoqin Zhang, Yizheng Wang, and Weiming Hu. "An Experimental Study on Exploring Strong Lightweight Vision Transformers via Masked Image Modeling Pre-training". In: *International Journal of Computer Vision* 133.7 (2025), pp. 3918–3950. DOI: 10.1007/s11263-024-02327-w.

[52] Haigen Hu, Xiaoyuan Wang, Yan Zhang, Qi Chen, and Qiu Guan. "A comprehensive survey on contrastive learning". In: *Neurocomputing* 610 (2024), p. 128645. DOI: 10.1016/j.neucom.2024.128645.

[53] Xin Gao, Zhihang Meng, Xin Jia, Jing Liu, Xinping Diao, Bing Xue, Zijian Huang, and Kangsheng Li. "An imbalanced binary classification method based on contrastive learning using multi-label confidence comparisons within sample-neighbors pair". In: *Neurocomputing* 517 (2023), pp. 148–164.

[54] Asif Newaz and Farhan Shahriyar Haq. "A novel hybrid sampling framework for imbalanced learning". In: *arXiv preprint arXiv:2208.09619* (2022).

[55] Maxime Oquab et al. "DINOV2: Learning robust visual features without supervision". In: *arXiv preprint arXiv:2304.07193* (2023). DOI: 10.48550/arXiv.2304.07193.

[56] Samer Albahra, Tom Gorbett, Scott Robertson, Giana D'Aleo, Sushasree Vasudevan Suseel Kumar, Samuel Ockunzzi, Daniel Lallo, Bo Hu, and Hooman H. Rashidi. *Artificial intelligence and machine learning overview in pathology laboratory medicine: A general review of data preprocessing and basic supervised concepts*. 2023. DOI: 10.1053/j.semdp.2023.02.002.

[57] Mauro Andrés Nievas Offidani and Claudio Augusto Delrieux. "Dataset of clinical cases, images, image labels and captions from open access case reports from PubMed Central (1990–2023)". In: *Data in Brief* 52 (2024), p. 110008.

[58] Philipp Thölke, Yorguin-Jose Mantilla-Ramos, Hamza Abdelhedi, Charlotte Maschke, Arthur Dehgan, Yann Harel, Anirudha Kemtur, Loubna Mekki Berrada, Myriam Sahraoui, Tammy Young, et al. "Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data". In: *NeuroImage* 277 (2023), p. 120253.

[59] Mateusz Lango and Jerzy Stefanowski. "What makes multi-class imbalanced problems difficult? An experimental study". In: *Expert Systems with Applications* 199 (2022), p. 116962.

[60] Krystyna Napierala and Jerzy Stefanowski. "Types of minority class examples and their influence on learning classifiers from imbalanced data". In: *Journal of Intelligent Information Systems* 46.3 (2016), pp. 563–597.

[61] K. Rajendran, M. Jayabalan, and V. Thiruchelvam. "Predicting breast cancer via supervised machine learning methods on class imbalanced data". In: *International Journal of Advanced Computer Science and Applications (IJACSA)* 11.8 (2020), pp. 57–61. DOI: 10.14569/IJACSA.2020.0110808.

[62] K. Hu, Y. Huang, W. Huang, H. Tan, Z. Chen, Z. Zhong, X. Gao, et al. "Deep Supervised Learning Using Self-Adaptive Auxiliary Loss for COVID-19 Diagnosis from Imbalanced CT Images". In: *Neurocomputing* 458 (2021), pp. 232–245. DOI: 10.1016/j.neucom.2021.06.047.

[63] Carmen Esposito, Gregory A Landrum, Nadine Schneider, Nikolaus Stiefl, and Sereina Riniker. "GHOST: adjusting the decision threshold to handle imbalanced

data in machine learning". In: *Journal of Chemical Information and Modeling* 61.6 (2021), pp. 2623–2640. DOI: 10.1021/acs.jcim.1c00160.

[64] Amjad Rehman, Teg Alam, Muhammad Mujahid, Faten S Alamri, Bayan Al Ghofaily, and Tanzila Saba. "RDET stacking classifier: a novel machine learning based approach for stroke prediction using imbalance data". In: *PeerJ Computer Science* 9 (2023), e1684. DOI: 10.7717/peerj-cs.1684.

[65] Devashish Pandey, Ayush Goyal, and Preeti Nagrath. "Class imbalance mitigation in predictive modelling of chronic diseases: a cost-sensitive learning framework". In: *Multimedia Tools and Applications* (2024). DOI: 10.1007/s11042-024-19705-8.

[66] Wendi Qu, Indranil Balki, Mauro Mendez, John Valen, Jacob Levman, and Pascal N. Tyrrell. "Assessing and mitigating the effects of class imbalance in machine learning with application to X-ray imaging". In: *International Journal of Computer Assisted Radiology and Surgery* 15.12 (2020), pp. 2041–2048. DOI: 10.1007/s11548-020-02260-6.

[67] M. Elbatel, H. Wang, R. Mart, H. Fu, and X. Li. "Federated Model Aggregation via Self-Supervised Priors for Highly Imbalanced Medical Image Classification". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (Oct. 2023), pp. 334–346. DOI: 10.1007/978-3-031-47401-9_32.

[68] Q. Su, H. N. A. Hamed, M. A. Isa, X. Hao, and X. Dai. "A GAN-Based Data Augmentation Method for Imbalanced Multi-Class Skin Lesion Classification". In: *IEEE Access* 12 (2024), pp. 16498–16513. DOI: 10.1109/ACCESS.2024.3360215.

[69] Yutong Li, Dong Wei, Jun Chen, Shuai Cao, Hongwei Zhou, Yong Zhu, and Yefeng Zheng. "Efficient and effective training of COVID-19 classification networks with self-supervised dual-track learning to rank". In: *IEEE Journal of Biomedical and Health Informatics* 24.10 (2020), pp. 2787–2797. DOI: 10.1109/JBHI.2020.3018181.

[70] Zhixian Liu, Qingfeng Chen, Wei Lan, Huihui Lu, and Shichao Zhang. "SSLDTI: A novel method for drug-target interaction prediction based on self-supervised learning". In: *Artificial Intelligence in Medicine* 147 (2024), p. 102778. DOI: 10.1016/j.artmed.2024.102778.

[71] Wen Chen and Ke Li. "Self-supervised Learning for Medical Image Classification Using Imbalanced Training Data". In: *Exploration of Novel Intelligent Optimization Algorithms*. Ed. by Ke Li, Ying Liu, and Wei Wang. Vol. 1590. Communications in Computer and Information Science. Springer, Singapore, 2022, pp. 255–265. DOI: 10.1007/978-981-19-4109-2_23.

[72] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. "Big self-supervised models advance medical image classification". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 3478–3488.

[73] Ghazaleh Mirzaee, Gianfranco Doretto, and Donald Adjeroh. "Multi-label Classification using Self-Supervised Learning: Addressing Class Inter-Dependency and Data Imbalance". In: *2024 International Conference on Machine Learning and Applications (ICMLA)*. 2024, pp. 1271–1276. DOI: 10.1109/ICMLA61862.2024.00198.

[74] Zhen Xu, Dinggang Lu, Jing Luo, Yang Wang, Jian Yan, Ke Ma, Tao Zhou, Xing Li, and Ren Tong. "Anti-interference from noisy labels: Mean-teacher-assisted confident learning for medical image segmentation". In: *IEEE Transactions on Medical Imaging* 41.11 (2022), pp. 3062–3073. DOI: 10.1109/TMI.2022.3187981.

[75] Shuquan Ye, Yan Xu, Dongdong Chen, Songfang Han, and Jing Liao. "Learning a Single Network for Robust Medical Image Segmentation with Noisy Labels". In: *IEEE Transactions on Medical Imaging* 43.9 (2024), pp. 3188–3199. DOI: 10.1109/TMI.2024.3389776.

[76] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. "Segment anything in medical images". In: *Nature Communications* 15.1 (2024), p. 654. DOI: 10.1038/s41467-024-44824-z.

[77] Minqing Zhang, Jiantao Gao, Zhen Lyu, Weibing Zhao, Qin Wang, Weizhen Ding, Sheng Wang, Zhen Li, and Shuguang Cui. "Characterizing Label Errors: Confident Learning for Noisy-Labeled Image Segmentation". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Ed. by Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin

Zhou, and Leo Joskowicz. Vol. 12261. Lecture Notes in Computer Science. Springer, 2020, pp. 721–730. DOI: 10.1007/978-3-030-59710-8_70.

[78] Banafshe Felfeliyan, Abhilash Hareendranathan, Gregor Kuntze, Stephanie Wichuk, Nils D Forkert, Jacob L Jaremko, and Janet L Ronsky. "Weakly Supervised Medical Image Segmentation with Soft Labels and Noise Robust Loss". In: *International Conference on Pattern Recognition*. Springer Nature Switzerland, Aug. 2022, pp. 603–617. DOI: 10.1038/s41467-022-28818-3.

[79] Shangkun Liu, Yanxin Li, Qing-wei Chai, and Weimin Zheng. "Region-scalable fitting-assisted medical image segmentation with noisy labels". In: *Expert Systems with Applications* 238 (2024), p. 121926. DOI: 10.1016/j.eswa.2023.121926.

[80] Haoyang Wu, Huan Wang, Hao He, Zixiao He, and Guotai Wang. "A novel weakly supervised framework based on noisy-label learning for medical image segmentation". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. Nice, France, 2021, pp. 1768–1772. DOI: 10.1109/ISBI48211.2021.9433883.

[81] Jialin Shi and Ji Wu. "Distilling effective supervision for robust medical image segmentation with noisy labels". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2021: 24th International Conference*. Springer International Publishing. 2021, pp. 668–677.

[82] Xiaodan Li, Yujia Wei, Qiang Hu, Chunfeng Wang, and Jian Yang. "Learning to Segment Subcortical Structures from Noisy Annotations with a Novel Uncertainty-Reliability Aware Learning Framework". In: *Computers in Biology and Medicine* 151 (2022), p. 106326. DOI: 10.1016/j.compbiomed.2022.106326.

[83] Erjian Guo, Zicheng Wang, Zhen Zhao, and Luping Zhou. "Imbalanced Medical Image Segmentation With Pixel-Dependent Noisy Labels". In: *IEEE Transactions on Medical Imaging* 44.5 (2025), pp. 2016–2027. DOI: 10.1109/TMI.2024.3524253.

[84] Hui Wang, Sos Agaian, and Desheng Liu. "A Noise Label Correction Architecture for Medical Image Segmentation". In: *2024 IEEE 3rd Industrial Electronics Society Annual On-Line Conference (ONCON)*. 2024, pp. 1–6. DOI: 10.1109/ONCON62778.2024.10931569.

[85] Coby Penso, Lior Frenkel, and Jacob Goldberger. "Confidence Calibration of a Medical Imaging Classification System That is Robust to Label Noise". In: *IEEE Transactions on Medical Imaging* 43.6 (2024), pp. 2050–2060. DOI: 10.1109/TMI.2024.3353762.

[86] Shuo Yang, Erkun Yang, Bo Han, Yang Liu, Min Xu, Gang Niu, and Tongliang Liu. "Estimating instance-dependent Bayes-label transition matrix using a deep neural network". In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2022, pp. 25302–25312.

[87] De Cheng, Tongliang Liu, Yixiong Ning, Nannan Wang, Bo Han, Gang Niu, Xinbo Gao, and Masashi Sugiyama. "Instance-dependent label-noise learning with manifold-regularized transition matrix estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 16609–16618.

[88] Zhaowei Zhu, Tongliang Liu, and Yang Liu. "A second-order approach to learning with instance-dependent label noise". In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. June 2021, pp. 10108–10118.

[89] Shikun Li, Xiaobo Xia, Jiankang Deng, Shiming Ge, and Tongliang Liu. "Transferring Annotator- and Instance-Dependent Transition Matrix for Learning From Crowds". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.11 (2024), pp. 7377–7391. DOI: 10.1109/TPAMI.2024.3388209.

[90] Yuan Wang, Huaxin Pang, Ying Qin, Shikui Wei, and Yao Zhao. "Adaptive estimation of instance-dependent noise transition matrix for learning with instance-dependent label noise". In: *Neural Networks* 188 (2025), p. 107464. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2025.107464.

[91] Ke Zou, Zhihao Chen, Xuedong Yuan, Xiaojing Shen, Meng Wang, and Huazhu Fu. "A review of uncertainty estimation and its application in medical imaging". In: *Meta-Radiology* 1.1 (2023), p. 100003.

[92] A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur. "Confidence calibration and predictive uncertainty estimation for deep medical image

segmentation". In: *IEEE Transactions on Medical Imaging* 39.12 (2020), pp. 3868–3878. DOI: 10.1109/TMI.2020.3006437.

[93] Thierry Judge, Olivier Bernard, Maria Porumb, Agisilaos Chartsias, Arjan Beqiri, and Pierre-Marc Jodoin. "Crisp-reliable uncertainty estimation for medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer Nature Switzerland. 2022, pp. 492–502.

[94] Lie Ju, Xin Wang, Lin Wang, Debashis Mahapatra, Xiaodong Zhao, Qibin Zhou, et al. "Improving medical images classification with label noise using dual-uncertainty estimation". In: *IEEE Transactions on Medical Imaging* 41.6 (2022), pp. 1533–1546. DOI: 10.1109/TMI.2022.3141425.

[95] Zeinab Sherkatghanad, Moloud Abdar, Mohammadreza Bakhtyari, Paweł Pławiak, and Vladimir Makarenkov. "BayTTA: Uncertainty-aware medical image classification with optimized test-time augmentation using Bayesian model averaging". In: *Knowledge-Based Systems* 327 (2025), p. 114123. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2025.114123.

[96] Tao Zhou, Yi Zhou, Guangyu Li, Geng Chen, and Jianbing Shen. "Uncertainty-Aware Hierarchical Aggregation Network for Medical Image Segmentation". In: *IEEE Transactions on Circuits and Systems for Video Technology* 34.8 (2024), pp. 7440–7453. DOI: 10.1109/TCSVT.2024.3370685.

[97] Figshare. *Brain Tumor Dataset*. https://figshare.com/articles/braintumordataset/1512427. Dataset.

[98] Shuailin Li, Zhitong Gao, and Xuming He. "Superpixel-guided iterative learning from noisy labels for medical image segmentation". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th Int. Conf.* Strasbourg, France: Springer International Publishing, Sept. 2021, pp. 525–535. DOI: 10.1007/978-3-030-87193-2_50.

[99] Carlos E. Cardenas et al. "Deep learning algorithm for auto-delineation of high-risk oropharyngeal clinical target volumes with built-in dice similarity coefficient

parameter optimization function". In: *International Journal of Radiation Oncology* Biology* Physics* 101.2 (2018), pp. 468–478. DOI: 10.1016/j.ijrobp.2018.01.114.

[100] Masoud Nickparvar. *Brain Tumor MRI Dataset*. https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset. Dataset.

[101] Mohamed Hany. *Chest CT scan Images*. https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images.

[102] Jie Zhang, Xiaohui Jia, Jingjing Zhou, Jin Zhang, and Jing Hu. "Weakly supervised solar panel mapping via uncertainty adjusted label transition in aerial images". In: *IEEE Transactions on Image Processing* 33 (2024), pp. 881–896. DOI: 10.1109/TIP.2023.3336170.

[103] Atika Akter and Sabbir Ahmed. *Brain Tumor Segmentation Dataset*. https://www.kaggle.com/datasets/atikaakter11/brain-tumor-segmentation-dataset. Dataset.

[104] Huan Minh Luu and Sung-Hong Park. "Extending nn-UNet for brain tumor segmentation". In: *Proc. Int. MICCAI Brainlesion Workshop*. Springer, 2021, pp. 173–186. DOI: 10.1007/978-3-031-09002-8_16.

[105] Alvaro Gonzalez-Jimenez, Simone Lionetti, Philippe Gottfrois, Fabian Gröger, Marc Pouly, and Alexander A. Navarini. "Robust t-loss for medical image segmentation". In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Cham, Switzerland: Springer Nature Switzerland, Oct. 2023, pp. 714–724. DOI: 10.1007/978-3-031-43898-1_68.

[106] Jiahui Yao, Yujie Zhang, Siyuan Zheng, Maitreyee Goswami, Prateek Prasanna, and Chao Chen. "Learning to segment from noisy annotations: A spatial correction approach". In: *arXiv preprint arXiv:2308.02498* (2023). DOI: 10.48550/arXiv.2308.02498.

[107] Hamid Rezatofighi, Nathan Tsoi, Jinwoo Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. "Generalized intersection over union: A metric and a loss for bounding box regression". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 658–666.

[108] Tianwei Zhang, Lequan Yu, Nanqing Hu, Sen Lv, and Shuxin Gu. "Robust Medical Image Segmentation from Non-Expert Annotations with Tri-Network". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2020*. Vol. 12261. Lecture Notes in Computer Science. Lima, Peru: Springer, Oct. 2020, pp. 249–258. DOI: 10.1007/978-3-030-59719-1_25.