Uncertainty-Aware Federated Dual-Branch Multimodal Fusion for Privacy-Preserving Skin Lesion Diagnosis

by

Ryan Azim Roup
22101669
Sabrina Sultana
21101045
Saba Tahsin
22101245
Mostakim Morshed
24141234
Marzia Khanam
21201571

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
June 2025.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Ryan Azim Roup

22101669

Sabrina Sultana

21101045

Saba Tahsin

22101245

Mostakim Morshed

Mostakim Morshed

24141234

Marzia Khanam

Marzia Khanam

21201571

Approval

The thesis titled "Uncertainty-Aware Federated Dual-Branch Multimodal Fusion for Privacy-Preserving Skin Lesion Diagnosis" submitted by

1. Ryan Azim Roup (22101669)
2. Sabrina Sultana (21101045)
3. Saba Tahsin (22101245)
4. Mostakim Morshed (24141234)
5. Marzia Khanam (21201571)

of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in Spring 2025.

Examining Committee:

Supervisor:
(Member)

Saiful

Mr. Md. Saiful Bari Siddiqui

Senior Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Mr. Dr. Farig Yousuf Sadeque

Associate Professor
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Associate Professor; Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

Skin lesion diagnosis is well-suited to data-driven methods, yet its real-world use is often hindered by strict privacy rules, fragmented data ownership, and inconsistent imaging conditions. To overcome these barriers, our work introduces a robust privacy-preserving framework that combines multimodal representation learning with decentralized training which reduces the need to centralize sensitive clinical data. We propose a unified dual-branch framework that supports (i) task-level multitask learning by coupling lesion classification with lesion segmentation using a shared Swin Transformer encoder and an Attention U-Net decoder, and (ii) modality-level learning by fusing image features with auxiliary patient metadata through configurable fusion strategies. To address client heterogeneity under federated learning, we introduce a Dynamic Uncertainty-Aware Divide2Conquer (DUA-D2C) aggregation strategy that adaptively weights client updates based on predictive performance and uncertainty, improving robustness in non-IID settings. Extensive evaluations on HAM10000 and MILK10K demonstrate that multimodal fusion improves over image-only baselines, and that DUA-D2C preserves or improves performance on HAM10000 while constraining degradation on the more heterogeneous MILK10K setting.

Keywords: Skin Disease Detection, Skin Cancer, Melanoma, Deep Learning, HAM10000, MILK10K, U-net, Attention U-net, Swin Transformer, DeepLabV3,Federated Learning,Multimodal Fusion,Uncertainty-Aware Aggregation, Dual-Task Learning

Acknowledgement

We express our sincere gratitude to the Almighty for His guidance and blessings throughout the completion of this research. We are deeply thankful to our supervisor, Mr. Md. Saiful Bari Siddiqui, for his valuable guidance, insightful feedback, and continuous support, which were instrumental in shaping this work. We also extend our sincere appreciation to our co-supervisor, Dr. Farig Yousuf Sadeque, for his constructive suggestions, encouragement, and motivation that significantly enhanced the quality of this research.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Skin diseases affect more than 1.9 billion people worldwide and represent one of the most common reasons for medical consultations, accounting for a significant proportion of visits to dermatology clinics [2]. Despite their high prevalence, accurate diagnosis of skin conditions remains challenging due to the large visual similarity between different lesion types and the subtle appearance of early-stage malignancies. Even experienced dermatologists may face difficulty in distinguishing benign from malignant lesions, particularly under time constraints or limited clinical resources.

Dermatoscopic and clinical imaging have become essential tools for supporting dermatological diagnosis. However, manual interpretation of these images is highly subjective and depends heavily on clinical expertise. Variations in image quality, illumination, skin tone, and lesion morphology further complicate the diagnostic process, increasing the risk of misclassification and delayed treatment, especially for aggressive conditions such as melanoma.

Recent advances in deep learning have demonstrated significant potential in automated skin lesion analysis. Convolutional Neural Networks (CNNs) and transformer-based architectures have achieved promising performance in tasks such as lesion classification and segmentation, often approaching or exceeding expert-level accuracy in controlled settings. Publicly available datasets have played a key role in accelerating research by providing large-scale annotated images for model training and evaluation.

Despite these advances, several challenges limit the real-world deployment of deep learning models in clinical environments. Many existing approaches focus on single-task learning, neglecting the complementary relationship between lesion localization and diagnosis. Additionally, most models rely on centralized training, which requires aggregating sensitive patient data in a single location, raising concerns related to privacy, data governance, and institutional regulations.

Furthermore, dermatological diagnosis often benefits from contextual clinical information, such as patient demographics and lesion history, which is rarely incorporated into image-only models. The absence of such multimodal integration can limit generalization across diverse patient populations and clinical settings.

These challenges highlight the need for diagnostic frameworks that jointly address task interdependence, data heterogeneity, and privacy constraints. Hence, our thesis is motivated by the development of a unified deep learning approach that combines multi-task learning, multimodal data fusion, and privacy-preserving training strategies to support reliable and scalable dermatological diagnosis.

1.2 Rationale of the Study

One of the most common and even fatal illnesses is skin cancer. The entire world is where early and precise diagnosis is important in patient prognosis. Deep learning computer-aided diagnosis systems have in recent years produced results, providing good performance in skin lesion analysis in controlled experimental environments. They are, however, difficult to implement in the real world due to data privacy issues, nonhomogeneous clinical settings, and the inaccessibility of annotated medical data. The majority of diagnostic systems have been developed based on centralized learning paradigms, which demand combining sensitive information about patients across different institutions. Such approaches raise serious privacy, security, and regulatory implications, and are impractical to support large-scale clinical collaboration. In addition, numerous studies pay their main attention to image-only classification, which may miss out on pathologically complementary clinical metadata and pixel-level lesion data that can strengthen diagnostic strengths and interpretability. Federated learning has come out as a viable alternative, as it facilitates collaboration. Direct data sharing is not required in model training. However, most federated aggregation algorithms, including Federated Averaging (FedAvg) [32], are very sensitive to non-IID data distributions, unreliable clients, and prediction uncertainty. These are also typical of multi-institutional medical datasets. These limitations can lead to volatile convergence and non-optimal world models. Inspired by these issues, this study examines an uncertainty-mindful federated learning system that interchangeably exploits multimodal knowledge, task-level adaptive client weighting, and joint supervision. This is specifically done by including predictive uncertainty in the aggregation process. The proposed approach is associated with data privacy, generalization, as well as clinical reliability and robustness.

1.3 Problem Statement

Despite notable progress in deep learning for skin lesion diagnosis, several critical challenges remain unresolved:

- Privacy constraints limit the feasibility of centralized learning across multiple medical institutions.
- Data heterogeneity across clinical sites leads to unstable federated training and biased global models.
- Limited utilization of multimodal information, such as clinical metadata and lesion segmentation, reduces diagnostic robustness.

- Naïve federated aggregation strategies fail to account for prediction uncertainty and unreliable client updates.
- Generalization across different imaging modalities, including clinical and dermoscopic images, remains insufficiently explored in a unified framework.

Therefore, there is a need for a privacy-preserving, uncertainty-aware federated learning framework that can effectively integrate multimodal information and handle heterogeneous medical data for reliable skin lesion diagnosis.

1.4 Objectives

The primary objective of this research is to design and evaluate a robust, privacy-preserving framework for skin lesion diagnosis using uncertainty-aware federated multimodal learning.

The key contributions of this thesis are summarized as follows:

- Design of a unified dual-branch learning framework that supports image-only, multimodal, and multi-task learning configurations for skin lesion analysis.
- Integration of Swin Transformer[19]–based visual encoders for hierarchical and robust feature extraction from dermoscopic and clinical images.
- Incorporation of an Attention U-Net[12]–based segmentation branch as an auxiliary task to enhance spatial awareness and representation learning when pixel-level annotations are available.
- Development of an uncertainty-aware federated aggregation strategy, termed DUA-D2C [49], that dynamically weights client updates using both predictive performance and uncertainty.
- Comprehensive evaluation across two heterogeneous datasets (HAM10000 [14] and MILK10K [48]), demonstrating improved robustness and generalization across imaging modalities.
- Systematic analysis of multimodal fusion strategies under both centralized and federated learning settings.
- Empirical demonstration of improved federated performance over standard FedAvg, particularly under client heterogeneity and uncertain predictions.

1.5 Methodology in Brief

The research methodology is concerned with the solution of the primary issues of dermatological diagnosis, including class imbalance, data heterogeneity, and the lack of data interchange among clinical institutes. Multi-task learning is included in the proposed strategy. Learning, multimodal data fusion, and privacy-preserving training paradigms will be developed in order to create a scalable diagnostic framework. This section provides an informative overview of the methodology pipeline that will be involved in the study.

The key components of the proposed methodology are summarized as follows:

- **Dataset Utilization:** The framework is tested on publicly available datasets of dermatological data in various imaging environments and with different distributions of classes.
- **Multi-Task Learning:** This is a dual-branch methodology of learning lesion classification and lesion segmentation that enables joint sharing of the complementary representations of tasks.
- **Shared Feature Encoding:** A visual feature encoder is shared in order to learn high-level representations, and task-specific heads are trained independently in order to attain classification and segmentation tasks.
- **Multimodal Fusion:** It can also be expanded to adopt additional clinical metadata on top of images in the form of Multimodal Fusion to enhance diagnosis.
- **Training Paradigms:** To capture the reality of data governance constraints in the real world, the proposed models are trained and assessed in centralized and federated learning settings.
- **Privacy-Preserving Learning:** Federated learning is utilized, allowing joint training of the model across distributed clients without revealing sensitive data of patients in a central place.
- **Uncertainty-Aware Aggregation:** The federated environment suggests an uncertainty-sensitive aggregation mechanism to ensure robustness under non-homogeneous distributions of client data.
- **Evaluation Strategy:** Model performance is evaluated using task-relevant measures of classification performance, segmentation performance, and stability of training in various learning configurations.

1.6 Scopes and Challenges

Scope of the Study

The study is aimed at designing and evaluating a multi-task learning, multi-modal feature fusion, and privacy-preserving training frameworks of a hybrid deep learning-based dermatological diagnosis framework. The paper concentrates on the power of the models having dissimilar data distributions and evaluates the model performance in the centralized and federated education situation by using publicly accessible datasets. The framework is not aimed at delivering formal compliance regulatory assurances despite its alignment with the principles of privacy-aware machine learning.

Future Scope

One possible direction of further research is the use of Large Language Models (LLM) to analyze the English-language unstructured clinical narratives, diagnostic notes, and medical reports. Such extensions may be useful in contextual reasoning and interpretability, though they need more datasets, evaluation processes and a strong emphasis on the privacy concerns, which is beyond the scope of the current study.

Challenges

Despite its potential, the proposed framework faces several challenges:

- **Data Heterogeneity:** The failure to generalize as a result of variations in image quality and acquisition circumstances, the availability or unavailability of some metadata in the datasets and institutions can restrict generalization.
- **Privacy–Performance Trade-off:** Privacy performance tradeoffs add overhead and instability in optimization in federated learning specifically.
- **Computational Complexity:** Multi-task learning, multi-modal fusion, and transformer based models increase the computational and memory requirements, which makes them impractical to apply.

Chapter 2

Literature Review

2.1 Preliminaries

This section introduces the key concepts and terminology that form the foundation of the literature reviewed in this chapter.

- **Skin Lesion Classification:** Image-based classification attempts to classify a dermatological image into diagnoses, i.e., benign or malignant lesions or disease classifications. This issue is generally presented as a multiclass classification problem, and is influenced by issues such as class imbalance, inter-class differences, and visual similarity between types of lesions.
- **Skin Lesion Segmentation:** This aims to identify the lesion region of an image. Clinically significant regions can be identified through correct segmentation, which can improve downstream classification and reduce background distortion. It is more commonly decided by overlap measures such as Dice score and Intersection over Union (IoU).
- **Multi-Task Learning:** Multi-task learning jointly optimizes classification and segmentation objectives within a single model. The method encourages application of complementary characteristics and has been shown to improve the strength and generalization during medical image analysis.
- **Multimodal Learning:** Multimodal learning uses visual information, like secondary clinical metadata, e.g., patient demographics or lesion properties. Multimodal models will perform better in terms of the generalization and diagnostic activity with the combination of context and images.
- **Federated Learning:** Federated learning is a decentralized training paradigm, which enables joint model training using distributed sources of data without exposing the raw patient data. Although it preserves privacy during learning, federated optimization faces challenges such as data heterogeneity, communication overhead, and training stability.

2.2 Review of Existing Research

Yu et al. (2021) [21] detect IoT-enabled remote skin disease using the proposed Targeted Ensemble Machine Classify Model (TEMCM) which combines multiple deep learning models to enhance classification accuracy. This research evaluates both traditional machine learning models and deep learning approaches such as K-Nearest Neighbors (KNN),Support Vector Machines (SVM), Artificial Neural Network (ANN) and Logistic Regression which were compared in earlier studies. However, this research focuses on Convolutional Neural Networks (CNNs) which includes custom models like Inception, ResNet50, Xception, MobileNet, DenseNet161 and VGG16. These models were tested under a combined evaluation framework to ensure fair comparisons and the preprocessing was done using techniques such as color featuring, ImageNet transfer learning, and data balancing. The main dataset that was used in this research was HAM10000 which consists of over 10,000 skin lesion images and covers 7 different seven skin conditions. There were other datasets as well in this research, such as the SD-198 (198 skin conditions) and the Skin Segmentation Dataset (SSD). These datasets were also analyzed however these datasets seemed less suitable for this study because they focus more on segmentation rather than classification. The data was split into three parts: 60% ,30% and 10% for training, testing and validation respectively. The 10% validation set contained images that the models had never seen before, which helped the researchers to test how well the models could work with new data. The Results suggest that DenseNet achieved the highest accuracy (86.5%) for all 7 skin conditions. On the other hand specialized models performed better in specific conditions for example, MobileNet performed exceptionally well in detecting Vascular Lesions (97% accuracy) but only has 68.4% accuracy for Melanoma. Hence, classification accuracy can be improved by choosing the most suitable model for each specific condition. So in this research the TEMCM approach combined all multiple deep learning models and achieved 98.48% accuracy.

Panneerselvam et al. (2025) [46] introduce a framework that integrates two neural network architectures. One model EfficientNetB3 is used for processing visual dermoscopic images and another model TabNet is used for analyzing clinical metadata such as patient age, gender, and medical history. The architectures are integrated together with an attention fusion mechanism which allows the model to focus on the most diagnostically relevant features from both modalities. This approach made the model better than the traditional ones and also focuses on overcoming the limitations of traditional single model methods (like CNN, DCNN) that rely more on image analysis. The methodology follows a structured process which is designed to maximize diagnostic accuracy. First, the dermoscopic images are first preprocessed through resizing, normalization, and data augmentation. On the other hand, the data is cleaned and then the EfficientNetB3 extracts high dimensional visual features from the images. In parallel, TabNet processes the clinical metadata using attention mechanisms to capture complex patterns. The results from both models were combined and passed into another neural network which was the final classifier. This network was trained for 50 rounds using the Adam optimizer (with 16 samples in each batch). The researchers used categorical cross-entropy as the loss function during training. Three major datasets were used in this study. First, ISIC 2018 and second, ISIC 2019, together include over 25,000 labeled dermoscopic images and third, HAM10000 with over 10,000 images that cover numerous types of skin disease. Five-fold cross-validation was used by the researchers to assess the correctness of the model. This multimodal achieved an surprising average accuracy of 98.69% on the ISIC 2018 dataset which surpasses all the previous single-modality approaches. The single modals typically reached around 92-96% accuracy. Furthermore, a detailed analysis showed that the model performed strongly in identifying melanoma cases with an accuracy of 98.3%, Actinic keratosis with accuracy 97.50%, Basal cell carcinoma with accuracy 98.40%, Melanocytic nevi with accuracy 98.75% , benign keratosis with accuracy 98.1% , Pigment lesion with 98.55% accuracy. Even though the multimodal achieved high accuracy, It came with a trade-off that is the training time increased by approximately 60% compared to the models that work with image only.

Wang et al. (2021) [27] propose an Interpretability-Based Multimodal Convolutional Neural Network (IM-CNN) for diagnosing skin lesions. The IM-CNN model has three parts: first, it handles patient metadata (such as age and lesion location). Second, extract domain knowledge-based features from segmented skin lesions (e.g., asymmetry, border irregularity, color variegation, and texture) and lastly to deal with skin lesion images. The methodology of IM-CNN can also be divided into three branches: first branch processes patient metadata. The second branch integrates image features with metadata and by using U-net, it isolates skin lesion areas and extracts image features, such as shapes and colors which mimics a dermatologist's perspective. First and the second branch are integrated together and processed with CatBoost classifier, this handles categorical and numerical features.The third branch uses EfficientNetB5 to capture the local and global details. To make it easier for dermatologists, interpretability modules are added to the third branch that consist of SHAP for metadata and Grad-CAM for skin lesion images. This study uses the HAM10000 dataset which consists of more than 10,000 dermoscopic images and over half the dataset consists of melanocytic nevus (NV) samples. So, to address this imbalance, IM-CNN model introduces focal loss (variant of the categorical cross-entropy loss) as the loss function. The dataset is split into 90% and 10% for training set and test set respectively using a stratified train-test split method which keeps the same class proportions in both parts. The results show that IM-CNN achieved an accuracy of 95.1%, sensitivity of 83.5%, specificity of 93.2%, and an AUC of 97.8 and this model performs better than models like DenseNet, ResNet, and Inception. Compared to single model performances, the IM-CNN model shows a 72% improvement in sensitivity and a 21% improvement in AUC_SEN_80 (this is a metric that focuses on high-sensitivity regions).

Hagerty et al. (2019) [15] present a hybrid approach for diagnosing melanoma from dermoscopy images. The approach combines deep learning with conventional image processing (handcrafted methods) to improve diagnostic accuracy. The two main components used in this study : a deep learning (DL) model and a handcrafted feature extraction model. The DL model components use a DL-based feature extraction using a pre-trained ResNet-50 network. The handcrafted methods have modules that detect medically relevant features such as atypical pigment networks (APN), color distribution patterns (via median color split algorithm), blood vessels (telangiectasia detection), and demographic information (age, gender, lesion location, etc.). These features are derived using biologically inspired algorithms and clinical data. The methodology starts with a few steps . First, the preprocessing steps such as hair and noise removal from dermoscopy images. Second, features were extracted by using both the handcrafted and deep learning methods. The outputs of individual modules from both approaches are then fused together using logistic regression, which combines their classification scores to produce a final melanoma probability prediction. Lastly , In order to improve the logistic regression model, researchers simplified the deep learning features using Principal Component Transform (PCT) which reduces the number of dimensions from the features. Two datasets were used in this study. The first consisted of 1,636 images which were collected from four clinics between the years 2007 to 2011 that contained 367 melanoma and 1,269 non-melanoma images. The second dataset was the HAM10000 dataset which contains 10,015 images but in order to create a binary classification, the researchers excluded nonmelanoma skin cancers and precancer. So finally the dataset was filtered to 9,174 images for binary classification (melanoma vs. non-melanoma). The results show that the ResNet-50 deep learning model alone achieved an AUC of 0.87, while the handcrafted model reached an AUC of 0.90. However, when both models were fused together, the AUC rose to 0.94 which shows that the fusion of deep learning and handcrafted features significantly improves diagnostic accuracy.

Yang et al. [42] address the difficulty of identifying skin disorders in places with restricted access to dermatologists in their study, "A Multimodal Approach to The Detection and Classification of Skin Diseases." In an effort to document how physicians use both images and patient data to establish diagnoses, they created a sizable new dataset by fusing over 37,000 skin photos with patient narratives produced by ChatGPT. Their multimodal method uses training tactics to increase accuracy and handle complex diagnostic decisions by combining robust picture classifiers with fine-tuned big language models. This method works well on consumer hardware and obtained a high accuracy of 91%, surpassing models that rely solely on text or graphics. In order to increase the usability and accessibility of AI-powered skin disease detection, the authors also intend to develop smartphone applications, incorporate real-time medical knowledge, and grow their dataset. As previously stated, there are possible problems with narratives produced by ChatGPT since they can never accurately replicate human interactions. Additionally, authors may adjust their training methods to artificially increase accuracy ratings.

To help with early skin disease diagnosis, Sumithra et al. [6] introduce a novel approach for automatically segmenting and classifying skin lesions. Using Gaussian filtering and morphological processes, they first preprocess skin pictures to eliminate noise and hairs. Then, they use an automatic region-growing segmentation algorithm to extract lesion sites. A complete collection of features is then extracted, totaling 4182 features per lesion. This includes 4096 RGB histogram features, 14 texture features from GLCM, and 72 color features from six color schemes. In classification, they employ k-Nearest Neighbors (k-NN) and Support Vector Machines (SVM), investigating both standalone classifiers and a fusion strategy with an In rule. Their system performed well in segmenting samples from Melanoma and Seborrheic Keratosis, but it had trouble with Bullae and Shingles because of the lower contrast between the skin and lesion areas. The system was tested on a custom dataset of 726 samples from five different disease classes: Melanoma, Bullae, Seborrheic Keratosis, Shingles, and Squamous cell. The fused classifier performed best in seborrheic keratosis, with the classification results peaking at an F-measure of 61.03% at a 70:30 train-test split.

Flosdorf et al. (2024) [36] address the significant skin cancer detection problem by applying Vision Transformers (ViT), a relatively recent deep learning architecture, to learn to classify images of skin lesions. From the Skin Cancer MNIST: HAM10000 dataset that has approximately 10,000 images across seven forms of skin cancer, the authors investigate the generalization performance of two larger sizes of a pre-trained ViT: ViT L16 and ViT L32. The main purpose of the study is to assess whether ViTs, especially their self-attention, provide better accuracy in skin lesion classification, with melanoma given a prominent stress owing to its high mortality. The findings indicate that ViT models perform better than conventional base models like decision tree classifiers (DTC) and K-nearest neighbor (KNN) classifiers, convolutional neural networks (CNNs), and smaller-sized models of ViT. Specifically, the ViT L16 model achieved an accuracy of 92.79% with a melanoma recall of 56.10%, while the ViT L32 model also achieved an accuracy of 91.57% with a melanoma recall of 58.54%. Moreover, ViT L16 and L32 also outperformed smaller ViT settings (ViT B16 and B32) in both accuracy and melanoma recall. The article highlights that the attention mechanism of the ViT identified the contours of lesions efficiently, which further attests to its superior performance. Indeed, the ViT L16 confusion matrix presented that a significant number of melanomas were falsely diagnosed as the most common melanocytic nevi. Certain limitations are noted. The uneven class distribution of the HAM10000 dataset might influence real-world generalizability. The melanoma recall of a little below 60% indicates that many of the actual melanoma cases go undetected, something that in itself poses a severe risk given the disease's high mortality rate. Secondly, ViTs, as "black-boxes" as most deep learning architectures, do not have the interpretability required for clinical adoption. Their high computational resource demands could also limit deployment in resource-limited healthcare settings. The use of data augmentation, in most cases beneficial for addressing class imbalance, runs the risk of overfitting when data augmented is not representative of actual variation in the real world.

Houssein et al. (2024) [38] presents a new Deep Convolutional Neural Network (DCNN) design for skin cancer lesion classification. This approach tackles the important problem of imbalanced datasets in medical images. The study used two imbalanced public datasets: HAM10000 (10,015 images, 7 classes) and ISIC-2019 (25,331 images, 9 classes). Images were preprocessed by normalization and resizing to 64x64 pixels. The method employed preprocessing (size reduction to 64x64 pixels, normalization) and random oversampling to tackle the imbalance of the dataset as initial model performance was low without it. The proposed DCNN architecture featured eight convolutional layers, followed by ReLU activations, batch normalization, pooling, and dense layers, ending in a softmax output for class probability. The model was trained with Adamax with categorical cross-entropy loss. The DCNN worked wonderfully. In the HAM10000 database, it achieved a test accuracy of 98.50%, along with impressive recall (98.51%), precision (98.56%), F1-score (98.48%), specificity (99.73%), and AUC (99.92%)71. Similarly, in the case of ISIC-2019 database, the DCNN achieved a good test accuracy of 97.11%, along with comparable high values for the other measures. The improved performance of the introduced DCNN was consistently evident compared to some of the state-of-the-art transfer learning models (VGG16, VGG19, DenseNet121, DenseNet201, MobileNetV2) and most of the earlier researches using the same datasets. Diagnostic accuracy was given utmost importance in the medical application, even though the DCNN required more computation and contained more parameters (68,326,247 parameters, 9.7 billion FLOPS). A limitation found was the absence of further advanced preprocessing stages for noise or artifact removal, suggesting directions for further development.This study proposes a novel DCNN structure for accurate skin cancer lesion classification, addressing the critical issue of imbalanced datasets within medical images. The study used two imbalanced public datasets: HAM10000 (10,015 images, 7 classes) and ISIC-2019 (25,331 images, 9 classes). Images were preprocessed by normalization and resizing to 64x64 pixels. The method employed preprocessing (size reduction to 64x64 pixels, normalization) and random oversampling to tackle the imbalance of the dataset as initial model performance was low without it. The proposed DCNN architecture featured eight convolutional layers, followed by ReLU activations, batch normalization, pooling, and dense layers, ending in a softmax output for class probability. The model was trained with Adamax with categorical cross-entropy loss. The DCNN worked wonderfully. In the HAM10000 database, it achieved a remarkable test accuracy of 98.50%, along with impressive recall (98.51%), precision (98.56%), F1-score (98.48%), specificity (99.73%), and AUC (99.92%)71. Similarly, in the case of ISIC-2019 database, the DCNN achieved a good test accuracy of 97.11%, along with comparable high values for the other measures. The improved performance of the introduced DCNN was consistently evident compared to some of the state-of-the-art transfer learning models (VGG16, VGG19, DenseNet121, DenseNet201, MobileNetV2) and most of the earlier researches using the same datasets. Diagnostic accuracy was given utmost importance in the medical application, even though the DCNN required more computation and contained more parameters (68,326,247 parameters, 9.7 billion FLOPS). A limitation found was the absence of further advanced preprocessing stages for noise or artifact removal, suggesting directions for further development.

Khullar et al. (2024) [44] aimed to recognize skin cancer types precisely with limited resources through a privacy-preserving federated transfer learning framework. The study utilized the ISIC-2019 dataset (25,331 images, 9 classes), which was preprocessed by resizing to 224x224 and 32x32 pixels, converting to grayscale, and augmenting data. They experimented with multiple lightweight pre-trained deep neural networks, such as EfficientNetV2S, EfficientNetB3, ResNet50, and NasNetMobile. One of the central contributions was the deployment of a Federated Learning (FL) platform, testing performance under both Independent Identically Distributed (IID) and Non-Independent Identically Distributed (Non-IID) data distributions between various clients. Baseline tests on 224x224 images revealed EfficientNetB3 to be one of the best performers with 87.86% accuracy (69.07% validation) before augmentation and 93.22% accuracy (90.33% validation) after augmentation. For low memory scenarios, downsampling images to 32x32 pixels with augmentation proved effective, with EfficientNetV2S yielding 91.67% accuracy and 92.36% validation precision with little loss. Under the federated learning setting, with the 32x32x3 augmented images, the validation accuracy was 89.83% with IID data and slightly higher at 90.64% with Non-IID data. The authors found that Non-IID distribution effect is not remarkable, which proves the model's robustness. The authors concluded that their method effectively reduced resource consumption without sacrificing the classification accuracy over some of the recent models. One of the most significant limitations encountered was the problem of applying this method in real time on handheld or mobile devices, a subject of future research.

Ahmadi Mehr and Ameri (2022) [25] study addresses the grave challenge of discerning skin cancers with accuracy, as it was the most common human neoplasm. Despite being the gold standard, old-fashioned visual inspection by specialists yields low accuracy, where experts achieve just around 60% for melanoma identification in the absence of any aids, while family doctors even less, i.e., 23-46%. Biopsy and histopathological investigations are confirmatory. The purpose of the research was to propose a viable deep learning (DL) method to support physicians, in this instance, by incorporating patient metadata (age, gender, and anatomical site) as well as lesion images as inputs. This was inferred from established relationships between these factors and skin cancer type and occurrence. As their method, the authors employed a pre-trained Inception-ResNet-v2 Convolutional Neural Network (CNN) that they picked for its good accuracy and computation efficiency in object recognition. The novelty was embedding patient metadata within the lesion image as a pseudo-QR-code rather than in the top-left corner pixels. This metadata was encoded as one-hot encoding for anatomical site and sex, and thermometer encoding for age, and an additional bit to encode image type (dermoscopic or photographic). The model was trained and tested on a large dataset of 66,735 clinical images from ISIC 2019/2020 Dermoscopic Archives, PAD-UFES-20, and Fitzpatrick17k databases with 16 skin conditions. There were two main classification tasks performed: binary (malignant or benign) and an 8-class classification of known conditions. Major findings indicated that patient metadata combination significantly enhanced classification accuracy, with the lowest accuracy gain being 5% in all scenarios considered. On the overall dataset, metadata-PB (Probability-based with metadata embedding) achieved 84.8% $\pm$ 0.9% for binary classification and 71.5% $\pm$ 1.8% for 8-class classification. Performance was significantly better when tested on a subset of 57,536 high-quality dermoscopic images: 94.5% $\pm$ 0.9% for binary and 89.3% $\pm$ 1.1% for 4-class classification. The metadata-PB approach also exhibited decreased confusion between malignant melanoma and benign melanocytic nevus. Limitations that were found include the technique's reduced impact on conditions that are represented mostly by lower-quality photo images (e.g., dermatitis, eczema, Kaposi sarcoma, and cutaneous lymphoma). Also, comparisons to other research were problematic due to differences in datasets and class numbers. Future research, the authors recommend, should strive to create bigger public datasets that have metadata as well as investigate sophisticated data augmentation techniques.

Tahir et al. (2023) [34] present the DSCC_Net, a new deep-learning-based convolutional neural network (CNN) model for the multi-classification of four types of skin cancer: melanoma (MEL), basal cell carcinoma (BCC), squamous cell carcinoma (SCC), and melanocytic nevi (MN). The authors stress the importance of early detection of skin cancer as the only way to ensure a good chance of survival, claiming that manual diagnosis is arduous. They test DSCC_Net on ISIC 2020, HAM10000, and DermIS, which are combined and curated especially for these four classes. Among the biggest hindrances in medical-based datasets is imbalance of classes. To mitigate this, the authors use up-sampling methods by generating synthetic samples for the minority classes using SMOTE-Tomek. This balances the distribution of the datasets. The DSCC_Net architecture consists of five convolutional blocks, ReLU activation, one dropout layer (0.2 rate) to prevent overfitting, and two dense layers ending with a SoftMax classification layer. The total trainable parameters of the network encompass 1,149,5, which indicates a relatively low complexity. The proposed DSCC_Net, by design and working principle, was carefully and exhaustively compared with ResNet-152, Vgg-16, Vgg-19, Inception-V3, EfficientNet-B0, and MobileNet. By applying SMOTE Tomek, DSCC_Net bagged outstanding results of an accuracy of 94.17%, precision of 94.28%, recall of 93.76%, F1-score of 93.93%, and an AUC of 99.43%. The performance easily surpasses the baseline models with MobileNet topping out in terms of accuracy at 92.51%. This shows the major advantage of SMOTE Tomek vis-a-vis not using it, for without USMT, DSCCNet would have performed miserable at just 83.20% in accuracy and 58.09% in F1-score. Another contribution of the study was facilitating the interpretability aspect by employing Grad-CAM heatmaps that visualize the model looking at the right areas of the skin for diagnosis. One of the major limitations of this study stems from the issue that the proposed DSCC_Net model is mainly designed for fair-skinned people given that the publicly available datasets used are primarily composed of images of lighter skin tones. This creates some concern regarding the generalizability and fairness of the model across different ethnicities. Future work will focus on integrating blockchain and federated learning with deep attention modules for more favorable results for skin cancer classification and broader skin infections, addressing current limitations.

Mateen et al. (2024)[39] propose a hybrid deep learning solution for the early and accurate diagnosis of melanoma. Recognizing the issue of distinguishing melanoma from benign lesions due to their identical symptoms, the study focuses on a two-class classification (malignant or benign). The solution integrates three significant components: U-Net, Inception-ResNet-v2, and a Vision Transformer (ViT). The main dataset used includes the ISIC challenge for 2020, which consists of 33,126 dermoscopic images. Data preprocessing consisted of resizing, centering, and extensive data augmentation employing all possible horizontal/vertical inversions, scale transformation, rotation, zoom magnification, brightness modifications, and so on, all with the aim of improving image quality and producing additional images (about 6562) based on the malignant class to maintain the class imbalance. Image segmentation using the U-Net architecture targeted the isolation of the relevant regions of interest (ROIs) from dermoscopic images to prevent false detections and focus the analysis on the affected areas. Now, the extracted segmented masks with the original images are pushed forward to the Inception-ResNet-v2 model along with deep features extraction for which residual connections are used in order to attain better accuracy and better generalization. Later, the Vision Transformer processes these features and refines spatial relationships via its self-attention mechanism for precise classification. The hybrid framework yielded state-of-the-art performances on the ISIC 2020 dataset, reporting an accuracy of 98.65%, sensitivity of 99.20%, and specificity of 98.03%. The model's consistency was further validated while conducting an ablation study for melanoma images in the HAM10000 dataset, producing 98.44% accuracy on unseen data. The ablation study also evidenced the high impact of the segmentation process that improved the performance in terms of precision, recall, and F1-score for benign and malignant lesion diagnosis compared to the diagnosis without segmentation. For example, segmentation improved the precision in malignant diagnosis by 0.83%, recall by 2.38%, and F1-score by 1.6%. The authors emphasize that the performance of the framework is credited to the top-down approach that combines different operations. The future scope of this work includes developing a smartphone application that will integrate the trained model for assisting in early diagnosis of melanoma.

Recognizing the significance of high false negative rates in AI skin cancer diagnosis, which are primarily a consequence of imbalanced medical datasets, Lyakhov et al. (2023)[30] introduce a new multimodal neural network system that specifically adjusts for such data imbalances in order to enhance recognition accuracy and reduce false negatives. For the first time, the system merges two very heterogeneous modalities of data: dermatological images and patient metadata, consisting of age, gender, and lesion localization. In terms of visual data, a preprocessing step occurs prior to the actual analysis, aiming to remove artifacts of hair structures that could subsequently interfere with diagnostically relevant features. Metadata is preprocessed using one-hot encoding. Another of the core novelties of the system is the adaptation of the cross-entropy loss function with weighting factors that increase the loss upon misclassification of minority (malignant) classes, which directly addresses the drawback arising from imbalanced datasets. For the analysis of visual data, the system uses some well-known pre-trained CNN architectures-DenseNet_161, Inception_v4, and ResNeXt_50-while metadata undergoes processing via an MLP, with their feature representations subsequently concatenated in a single concatenation layer. The system has been thoroughly trained and tested on 41,725 skin cases from the ISIC Archive in 10 clinically relevant categories. Highly promising results have been recorded. The DenseNet_161 has recorded the most accurate recognition with respect to the 10 diagnostic classes of 85.20%. This is a significant improvement with accuracy rates ranging from 1.99 to 4.28 percentage points higher than those of the original multimodal systems that did not use the modified loss function. Most importantly, the modified loss function led to a decline in the instances of false negatives; Inception_v4 suffered 234 fewer false negative cases, whereas DenseNet_161 and ResNeXt_50 were able to reduce 10 and 42 cases, respectively, hence better detection of malignant lesions. The DenseNet_161 is also statistically robust with a sensitivity of 0.8520, specificity of 0.9836, and a Matthews Correlation Coefficient (MCC) of 0.7168, which is found to be reliable enough for unbalanced datasets. The authors suggest that this overall improvement is as a result of the "synergy" applied in the combination of hair cleaning, heterogeneous data analysis, and the balanced learning approach. Despite its strengths, this system does present some critical limitations. Namely, a false negative prediction remains a risk, albeit a much smaller one, and this system cannot be used as a stand-alone diagnostic tool-the emphasis is placed on its use as a high-precision support tool for dermatologists. The model's capability of generalizing is limited for rare types of pigmented skin tumors or unusual localizations (e.g., palms), as these were not sufficiently included in training, so the diagnosis may be erroneous there. Another limitation is an inability to identify, classify, and evaluate two or more pigmented skin tumors from a single image; it just processes the lesion with the most contrast or largest lesion, thus lacking explicit detection among individual lesions. This weakness creates an opportunity for both missed diagnoses and incorrect classifications on images containing no lesions. The authors state that future work can look into more complicated ensemble systems along with incorporating semantic segmentation to solve these shortcomings.

Anjali Jain et al. (2021) [18] presented a comparative analysis of six deep learning transfer learning models for multi-class skin cancer classification from the HAM10000 dataset. The authors underline the difficulty of diagnosing skin cancer for a dermatologist because of the visual similarity of different skin cancer pigments, stressing the importance of early detection. Their work aims at classifying images of the seven categories of skin cancer in the HAM10000 dataset that is highly unbalanced to begin with. In order to treat the class imbalance in the dataset, their solution heavily involves data augmentation through image replication of low-frequency classes. This, in turn, increased the size of the training set, ensuring relatively equal amounts of images per class (e.g., Dermatofibroma images multiplied 110 times; 3080 is close to Melanocytic Nevi's 3179). Other augmentation factors were rotated, zoomed, and shifted vertically and horizontally. Preprocessing of the images involves normalization of pixel levels, followed by resizing to dimensions of 128*128. The study considers six pre-trained transfer learning models-VGG19, InceptionV3, InceptionResNetV2, ResNet50, Xception, and MobileNet-all initially trained on ImageNet weights and then retrained on the augmented HAM10000 dataset. Experimental results comparing performance with or without image repetition show replication is preferred for this setup. Among all models with data repetition, the studio Xception Net has achieved the best rates with 90.48% accuracy, 89.57% average recall, 88.76% average precision, and 89.02% average F-Measure. The lowest loss value, 0.5168, was also obtained by Xception Net. InceptionResNetV2 and MobileNet performed well next to it. Those authors state that "Akiec" (actinic keratosis, label 0) and "Mel" (melanoma, label 5) were the most difficult lesions to identify correctly, due to their extreme resemblance to simple skin patches; and, on the other hand, "Nv" (melanocytic nevi, label 4) was the skin lesion with the highest classification rate. Although the study considers class imbalance and selects the high-performing models, a limitation to this study is that certain models (VGG19, InceptionV3, ResNet50) showed a decrease in accuracies when repeated versus without repetition. Hence, the authors believe that more extensive research and proper fine-tuning of transfer learning algorithms would provide better accuracies.

Khalaf et al. (2025)[43] have presented a comprehensive review of the applications of deep learning (DL) in skin cancer diagnosis, especially considering developments in the fields of image segmentation and classification. The authors put forth that DL has been instrumental in enhancing the accuracy, efficiency, and timeliness of diagnosis for better outcome and early detection and treatment of patients. The paper surveys various DL-based models, mainly convolutional neural network (CNN) based, for skin lesion detection. For segmentation models, U-Net variants such as U-Net++, and U-Net3+, incorporation of nested skip paths, and feature aggregation, have been explored to enhance accuracy. Other segmentation networks include, but are not limited to, O-Net and Mask R-CNN, generally adopting an encoder-decoder layout. For classification purposes, it is reviewed in the paper how TL has been studied and applied by means of several architectures such as ResNet50, ResNet152, SqueezeNet1.1, and DenseNet121 with regard to its potential to enhance diagnostic accuracy, especially when data available are limited. The integration of attention mechanism within the segmentation networks such as RSUnet and multi-level attention models is said to increase the accuracy by focusing on the crucial features and especially where the boundaries of the lesions are irregular. Aside from XAI methods, the paper examined Grad-CAM and SHAP as these are important from an interpretation standpoint and trust perspectives of clinical settings. The most recent methods, including transformer-based architectures (ViTs, ISC-TransUNet, FAT-Net) and hybridization of transformers and CNNs, focus on capturing global dependencies and local features. Yet, these achievements notwithstanding, the authors try to underscore a few major challenges that slow down the clinical application of DL models. Dataset diversification, in particular, the underrepresentation of darker skin tones-which allow for a biased prediction of a model, thus limiting its generalizability-is one such concern. Model interpretability and trust weigh heavily on AI decision processes because doctors have to know. Practical matters relevant to real-world implementation, including the huge computational complexity in DL models, adverse environments with minimal resources, high-cost infrastructure, etc., are also considered. In their opinion, the major research topics are directed toward making more-efficient transformer-based methods, handling bias, and increasing classification so as to consider more kinds of skin conditions.

Gururaj et al. (2024) [28] have offered a deep learning method for early and accurate detection of skin cancer when treatment options are limited due to the advances of the disease. The study opts for CNNs because of their established superiority in object detection and classification. The research uses the HAM10000 dataset. An important feature of their approach is the data preprocessing mainly to accommodate the dataset imbalances and noise aspects. Sampling techniques such as oversampling and undersampling are adopted to maintain a better balanced distribution of classes for the HAM10000 dataset. To augment the reliability, it incorporates the famous "Dull Razor" technique that removes the hair artifacts in skin images, a common hurdle that usually hinders accurate diagnosis. The image segmentation process is carried out using an autoencoder-decoder architecture to segment out relevant regions of interest. For classification, transfer learning is used by the authors, relying on pre-trained models, specifically DenseNet169 and Resnet50. Experiments involved authors comparing the performance of these transfer learning models under different data sampling strategies and training-testing ratios. Some of the key results indicate that the DenseNet169 model in underscore sampling produced 91.2% accuracy and 91.7% F1-measure, whereas the Resnet50 model gave 83% accuracy and 84% F1 measure under oversampling. Therefore, the paper concludes that DenseNet169 with undersampling performed better on classifying the seven types of skin lesions in the HAM100000 dataset. An important limitation that was discovered during the study was that the DenseNet169 architecture does not allow the oversampling technique while the Resnet50 architecture does not allow the undersampling technique. The authors have suggested that future work should be targeted on increasing forecast accuracy through further parameter tuning. While their study propounds that their work yields better results than several previously published CNN models (claiming 4% higher than and 5% higher than from their reference list), it remains difficult to establish a direct numerical comparison, since the exact reported accuracy values for the reference models are not mentioned in the given excerpt of their work.

Wiest et al.[41] paid special attention to the problem of understanding information accurately while upholding the confidentiality of patient details in healthcare contexts involving large language models. A way to use LLMs that maintains privacy by querying structured EHR is presented in the study. The main advantage comes from using split processing which means that parts related to privacy are done locally, but the basic language part is handled by a faraway large language model. The system uses hybrid architecture to ensure that no personal information about patients is ever revealed to assist in query generation. The research shows positive outcomes in building structured queries while using the datasets from MIMIC-IV. While the strategy greatly lowers the risk of privacy breaches, it is still based on trust in nearby computing and on secure models for the API. Given that patient data is involved in clinical settings, more research is required on guarantees for privacy, mainly regarding differential privacy and deployments using federated data. Nevertheless, this approach matches the objectives of hybrid deep learning frameworks that care about performance and privacy which is especially significant for clinical support systems in dermatology and similar areas.

Luu et al. [29] develop a privacy-focused approach for skin disease classification, so it closely fits the goals of this thesis. The approach is built on federated learning, so several clients can join together to build a model without exchanging their raw data. DP mechanisms and secure group data processing are included in FL by the paper to increase both privacy and security. To test the model, the authors use real-world images and they find that the approach is accurate and ensures no patient information is revealed. The findings reveal that small amounts of DP noise do not seriously impact performance, so a decent balance can be reached between privacy and accuracy. Yet, people recognize that there are challenges such as extra time required for communication and difficulties with models reaching convergence. The paper directly helps our hybrid architecture in training useful dermatology models from different data sources without breaking data privacy laws or ethical rules.

Gupta et al. [37] introduced PP-HFL as part of the research in for assisting mental healthcare via clustering and quantum-inspired methods. Since its subject is different, this work is still very valuable for our thesis because it focuses on joint multi-institutional efforts that do not involve sharing data, just like what we plan to do in dermatological AI. By clustering clients by their behavior and fine-tuning the local models, the method boosts the accuracy of diagnostics and saves time in the whole process. Also, a new quantum-inspired technique is included to boost the model's convergence and privacy. Analysis of psychological datasets that require privacy protection indicates that PP-HFL provides better accuracy and achieves faster convergence than the usual federated learning. Thanks to this approach, we can come up with helpful solutions for our system such as adaptive clustering of medical institutions, smart privacy controls based on resources and potential use of quantum-inspired privacy-preserving optimization in deep learning for medical support systems. The privacy-enhancing structure and automated personalization in the paper might be useful in our dermatology system with multi-modal data, despite the original subject not being skin diseases.

Mahmud et al. [35] propose using SkinNet-14, a deep learning model that performs skin cancer classification using images that are not very detailed, i.e., low resolution dermoscopy. With a basic CCT transformer network, the framework is able to manage images with a size of 32×32 pixels. As a result of using a lighter version of Transformer, it can be deployed more easily in resources-limited places. Various image processing and transforming techniques used in the model make the input more effective and help balance out the number of types of medical images. SkinNet-14 obtained 97.85%, 96.00% and 98.14% accuracy on HAM10000, ISIC and PAD datasets and its training speed per round was 5–15 seconds. SkilNet-14 performs both more accurately and steadily than previous transfer learning models when fewer samples are used. Nevertheless, the solution does not take into account clinical metadata or ways to protect a patient's privacy which are key for actual clinical use. A feature like this may allow for better application of the model where privacy and different ways of monitoring are important.

Ou et al. [26] introduces a deep learning based multimodal fusion model of skin lesion diagnosis through the use of smartphone The article presents a clinical data and the metadata of the patient is collected, a new method of classification of skin lesions with the use of combination of smartphone collected clinical images and metadata of the patient. The authors suggest a deep learning framework, named as MMF Net that attaches to both image or metadata representation in the same modality (image, metadata) by intra modality self attention and the different modality (image, metadata) by inter modality cross attention. The dual attention strategy of the model is used to separate important characteristics in each modality and about the same time provide guidance to each modality through relations based on surroundings held by the other. The model has been trained and tested on the PAD-UFES-20 dataset containing 2,298 image patches of smartphones with up to 21 metadata features, including the age of a patient, the location of lesion, and the type of skin. The authors made a five-fold cross validation and compared their model to the existing fusion methods such as MetaNet or MetaBlock. The best performance was seen in MMF Net with an increase of accuracy (61.61, 76.8), balanced accuracy (65.11, 77.5), and AUC (0.901, 0.947) all at a high significance level of $p < 0.001$. Markedly, the AUC values of the classes were high specifically, melanoma (0.98), nevus (0.97), and actinic keratosis (0.95), but there was also a certain degree of confusion between the classes of basal cell carcinoma and squamous cell carcinoma, which is a factor that reflects clinical practice. Practical orientation of the study on smartphone based data has been considered one of its strengths due to the fact that it expands the possibility of the solution to be applied in a diverse clinical setting, as well as the careful application of attention mechanisms in the implementation of feature fusion. Besides, inclusion of ablation studies assist in the illustration of the significance of every component in the model. Nevertheless, certain issues also inhibit the work as it involves a narrow range of only six types of lesions as well as failure to validate it on external datasets, which can cast doubt on the ability to apply it to wider population and more challenging dermatological cases.On the whole, the study is significant in the area of AI based diagnosis of dermatology. It also highlights the importance of using both visual and non visual information on patients in order to achieve better classification results and provides a framework upon future research on mobile based skin health solutions. To deploy it practically, there should be additional research on larger batch size, the inclusion of additional lesion types, and demonstrating the project on a different demographical and clinical settings.

2.3 Summary of Key Findings

Discuss key results, patterns or trends, and implications.

Author	**Year Published**	**Results/Findings**
Yu et al. [21]	2021	Targeted Ensemble Machine Classify Model (TEMCM) combined multiple deep learning models for skin disease detection which achieved 98.48% accuracy. DenseNet performed overall best with (86.5% accuracy), also models like MobileNet worked great in specific conditions.
Panneerselvam et al. [46]	2025	A multimodal framework which combines EfficientNetB3 for images and TabNet for clinical data, and both are integrated via attention fusion. This approach achieved 98.69% accuracy on ISIC 2018, which outperforms single-model approaches.
Wang et al. [27]	2021	A multimodal model (IM-CNN) which combines metadata, domain-specific features, and images for skin lesion diagnosis. It uses interpretability tools (SHAP, Grad-CAM) and focal loss to address class imbalance, it achieved 95.1% accuracy and outperformed models like DenseNet and ResNet in sensitivity and AUC.
Gouda et al. [23]	2022	To classify skin tumors as benign or malignant, ESRGAN-enhanced images and deep learning models has been used. InceptionV3 achieved the highest accuracy (85.8%) which outperforms custom CNNs and prior models like VGG19 and AlexNet. Image enhancement and transfer learning significantly improved classification performance.
Hagerty et al. [15]	2019	A hybrid melanoma diagnosis method which combines deep learning (ResNet-50) and handcrafted image features. Fusing both approaches using logistic regression improved performance and achieved an AUC of 0.94 (this is higher than using either method alone).

Author	Year Published	Results/Findings
Lyakhov et al. [31]	2023	Using a special loss function the model was able to detect malignant cases more accurately. The DenseNet-161 model with 85.20% accuracy with low false negative rate. This means it was quite reliable at catching serious skin conditions
Mohan et al. [45]	2023	Transformer-based models like ViT, Swin, and DinoV2 used to classify skin diseases. DinoV2-Base reaching 96.48% accuracy and 0.9728 F1-score, beating traditional CNNs. With the help of GradCAM and SHAP, the models also offered clear explanations for their predictions. The method showed strong results across datasets, though challenges like overfitting and high processing needs remain.
Aghdam et al. [22]	2023	Attention Swin U-Net, a transformer-based model intended to enhance skin lesion segmentation. Addresses frequent issues such as low contrast and uneven lesion forms. Outperforming U-Net and other sophisticated models, the model produced remarkable results on the ISIC 2017 dataset, with 92.40% Dice score and 96.56% accuracy. It also saves money on training and is effective.
Nam et al. [33]	2023	Instead of collecting all the data in one place, local CNN models were trained on users' own devices, and only the model updates were shared. The HAM10000 dataset was expanded to more than 46,000 photos by using SMOTE to create synthetic samples in order to address the class imbalance. The accuracy of the model increased from 79% to 95% when the data was balanced.
Shah et al. [40]	2023	This paper describes about how Large Language Models (LLMs) can help in dermatopathology by creating detailed reports and also helping with diagnoses. It is still new and needs more data. Especially for rare cases. The authors think combining LLMs with image based AI could make diagnosis easier and faster in the future
Sumithra et al. [6].	2015	Their system got about 61% accuracy for classifying skin lesions. It worked well for some diseases but had trouble when the skin and lesion looked similar. They suggest using better data and newer methods to improve results.

Author	Year Published	Results/Findings
Yan et al. [50]	2024	They developed MAKE to better connect skin images with medical info. This helped improve diagnosis accuracy by around 7.5%. It outperformed other models but still needs more testing in real clinical settings. Also it requires a lot of computing power.
Pham et al. [47]	2025	Using multiple deep learning models and combining skin pictures with patient information, they achieved an accuracy of about 84%. It outperformed any single model. Although the results were poorer when patient information. Or attention layers were removed. The model has trouble with uncommon skin types and is rather complicated. Real world testing and more varied data are required.
Rezvantalab et al. [13]	2018	DenseNet201 achieved the highest overall performance with a macro AUC of 98.16% and micro AUC of 98.79%. For melanoma and BCC, CNNs outperformed dermatologists, scoring up to 94.4% and 99.3% AUC, respectively.
Flosdorf et al. [36]	2024	ViT L16 outperforms all models in accuracy (92.79%). ViT L32 has the highest melanoma recall (58.54%), meaning it's better at catching melanoma cases, which is critical due to its lethality. Both models outperform CNNs and smaller ViT models in all major metrics.
Houssein et al. [38]	2024	The proposed DCNN achieved 98.5% accuracy, 98.56% precision, and 99.92% AUC on the HAM10000 dataset, outperforming VGG16, DenseNet, and MobileNet models. Even with unbalanced data, it showed good multiclass classification performance.
Khullar et al. [44]	2025	The federated EfficientNetV2S model achieved 90.64% accuracy on Non-IID data and 89.83% on IID data using low-resolution (32×32) images. After augmentation, EfficientNetB3 reached a validation accuracy of 94.74% on full-size images.
Mehr et al. [25]	2022	The model achieved 94.5% accuracy in binary classification (benign vs. malignant) and 89.3% in 4-class classification using dermoscopic images. Patient metadata added improved classification performance by at least 5% for every task.

Author	**Year Published**	**Results/Findings**
Tahir et al. [34]	2023	CNN based multi-classification model DSCC_Net achieved an AUC of 99.43%, an accuracy of 94.17%, a recall rate of 93.76%, a precision of 94.28%, and an F1-score of 93.93% over the dataset via SMOTE Tomek up-sampling.
Mateen et al. [39]	2024	In the ISIC 2020 domain, the hybrid framework of U-Net, Inception-ResNet-v2, and Vision Transformer yielded an accuracy of 98.65%, sensitivity of 99.20%, and specificity of 98.03%. Ablation studies concluded that segmentation yielded a performance of up to 98.44% accuracy.
Lyakhov et al. [30]	2023	A multimodal neural system incorporating a modified cross-entropy loss function for data balancing was able to obtain 85.20% recognition accuracy for 10 diagnostic categories, with the DenseNet_161 performing best. Weighting factors further heavily reduced false negative predictions and increased overall accuracy by 1.99-4.28 pp.
Anjali Jain et al. [18]	2021	Six transfer learning models, namely VGG19, InceptionV3, InceptionResNetV2, ResNet50, Xception, and MobileNet., were evaluated for multi-class skin cancer classification on HAM10000 dataset. Xception Net, supported by data balancing, outperformed all models with an accuracy of 90.48%, along with strong recall, precision, and F-Measure.
Khalaf et al. [43]	2025	CNNs integrated with transformers gain an accuracy up to 97.48%, but that comes at the expense of greater complexity, rendering them unsuitable for real-world deployment. Segmentation and DL classifier hybrid models are thus more feasible and accurate alternatives.
Gururaj et al. [28]	2024	DenseNet169 and Resnet50 were employed with both undersampling and oversampling data preprocessing techniques to classify seven different types of skin cancer using dermoscopic images. With an undersampling technique, the DenseNet169 model achieved an accuracy of 91.2% and an F1-measure of 91.7% and Resnet50 model yielded an 83% accuracy and an 84% F1-measure when oversampling was used

Author	Year Published	Results/Findings
Wiest et al. [41]	2024	Presented a privacy-sensitive system based on split processing and structured EHR querries and LLMs. Private information processing is performed locally whereas the language model is executed remotely. Accomplished effective generation of queries on MIMIC-IV with hybrid architecture. Deals with the risks to privacy but needs additional study on the subject of differential privacy and federated deployments.
Luu et al. [29]	2023	Designed a differential privacy (DP) and securely aggregated federated learning model of skin diseases classification. Obtained high accuracies on real world papers and guaranteed that patient data does not leak. The effect (on performance) of DP noise was negligible but concerns lie with communication overhead and convergence.
Gupta et al. [37]	2024	Advanced the PP-HFL of mental healthcare on the basis of the clustered federated learning and quantum inspired approaches. Being related to psychology, it provides helpful concepts such as adaptive clustering, privacy preservation, accelerated convergence which are applicable to the domain of dermatological AI involving multi-institutional data.
Mahmud et al. [35]	2024	Proposed SkinNet-14, a lightweight CCT transformer-based classifier of skin cancer on low-resolution dermoscopy (32 32). Attained a 97.85% (HAM10000), 96% (ISIC) and 98.14% (PAD). It was better than previous models that have rapid training speeds but no support of clinical metadata or privacy features.
Ou et et al. [26]	2022	Obtained good results on community data sets with multi-scale features. Nonetheless, important concerns about the patient privacy are still not addressed, which is essential in real-life clinical practice.

Table 2.1: Summary of Key Findings

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

This thesis develops and evaluates an uncertainty-aware federated dual-branch multimodal learning framework for skin lesion analysis, supporting both classification and segmentation tasks under centralized and federated settings. The framework is designed to operate on heterogeneous medical imaging datasets, including dermoscopic, clinical, and multimodal inputs augmented with structured patient metadata.

All visual inputs, whether grayscale or RGB, are standardized to a spatial resolution of 224×224 pixels. The learning pipeline accommodates both single-label and multi-label classification objectives, as well as pixel-level segmentation supervision when available. The final architectural configuration was selected through iterative experimental validation, balancing predictive performance, generalization ability, and computational efficiency across diverse datasets.

The proposed framework employs Swin Transformer backbones as shared visual encoders, lightweight metadata encoders for structured clinical information, modular fusion heads for multimodal integration, and an uncertainty-aware federated aggregation mechanism (DUA-D2C [49]). This modular design allows selective activation of architectural components depending on dataset characteristics while maintaining a unified optimization strategy.

The software stack is implemented in Python using the PyTorch deep learning framework with CUDA acceleration. Additional libraries include `torchvision` for data preprocessing and augmentation, `timm` for pretrained model initialization, and standard scientific Python tools for evaluation, visualization, and experiment logging. Experiments are conducted in both local and cloud-based environments.

Local Hardware Configuration All local experiments are executed on the following system:

- Operating System: Windows 11
- CPU: Intel Core i7-14700K

- GPU: NVIDIA RTX 4080 Super (16 GB VRAM)
- RAM: 64 GB DDR5
- Storage: 1 TB SSD + 1 TB HDD

Cloud Environment To validate reproducibility and scalability, selected experiments are also conducted on cloud infrastructure:

- Platform: Kaggle
- GPU: Dual NVIDIA T4 GPUs (parallel usage)

Evaluation emphasizes both effectiveness and robustness. Core performance metrics include accuracy, macro-averaged F1 score, and AUC–ROC. For segmentation tasks, Dice coefficient and Intersection over Union (IoU) are additionally reported. In federated experiments, predictive entropy is computed to quantify model uncertainty and guide client weighting during aggregation.

To mitigate variance and reduce the risk of over-interpreting individual runs, experiments are conducted with fixed random seeds and validated across multiple data splits where applicable. Performance comparisons between centralized and federated learning, as well as between different fusion strategies, are analyzed consistently using the same evaluation protocol.

3.2 Societal Impact

Skin diseases are among the most common health issues across worldwide and it can severely affect quality of life. In serious cases such as melanoma, delaying diagnosis can be life-threatening and would often lead to higher treatment costs. This research has significant societal impact by applying deep learning for automated segmentation and classification of skin lesions. It enables earlier detection, which improves survival rates and reduces costs.

3.3 Environmental Impact

This research mainly involves computational experiments using deep learning models for medical image analysis. Though, computer aided methods do not directly produce physical waste like industrial or chemical processes, training large neural networks requires significant computational power, which consumes energy and contributes to carbon emissions depending on the energy source. To reduce this, the project emphasizes efficient model selection like freezing the encoders and early stopping techniques, which reduce training time and resource usage. The long-term environmental benefits will be positive, as deploying AI-based diagnostic tools in healthcare can reduce unnecessary hospital visits and redundant procedures which would lower the overall environmental footprint of healthcare delivery.

3.4 Ethical Issues

This research involves the use of medical imaging data and patient-related metadata, which provokes significant ethical issues of privacy, fairness, and responsible application of artificial intelligence. As the images of skin lesion can include sensitive visual data as well as clinical characteristics, high levels of data protection need to be applied throughout the whole research cycle. Any dataset utilized in the present study is of a publicly accessible and ethically verified source at which anonymization of the patient identities is realized and previously registered informed consent by the respective data providers.

Ethical responsibility in the federated learning environment is established even more because processed patient data do not leave local client devices at all. Rather, encrypted model updates only are exchanged with the central server, which minimizes the possibility of data leakage. Uncertainty-aware aggregation in addition reduces the effect of untrustworthy or noisy client information, reducing the risk of biased or hazardous model execution.

The other important ethical concern is associated with algorithm bias and impartiality. The datasets of skin lesions are skewed in terms of skin tones, types of lesions, and demographic data. To overcome this, the suggested framework uses macro averaged evaluation measures, class weighted loss functions, and uncertainty estimation schemes to minimize the overconfidence on under-represented classes and encourage fair model behavior.

From a professional responsibility perspective, the system is designed strictly as a decision-support tool and not an expert replacement of clinical acumen. It is the role of researchers and developers to make the research transparent, well-validated with clear communication of the limitations of the model, especially in high-risk medical uses. The framework will assist in making informed clinical decisions and also encourages the safe, ethical, and responsible implementation of predictions in practical medical settings by delivering interpretable and uncertainty-aware forecasts.

3.5 Standards

The study follows the established standards and best practices in the medical image analysis, machine learning experimentation, and software development. Image preprocessing adheres to conventional standards of normalization and resizing used in the computer vision research, making it compatible with the pretrained architectures like Swin Transformers.

Regarding machine learning, the experimental protocol satisfies the requirements of reproducibility, such as fixed random seeds, constant data splits, and standard evaluation measurements, such as accuracy, F1-score, AUC-ROC, Dice coefficient, and IoU. Experiments of federated learning are consistent with the emerging privacy-protective AI standards because of a lack of centralized data storage and integration of uncertainty-aware aggregation techniques.

3.6 Project Management Plan

It implemented a systematic and incremental development strategy to complete the project on time with effective use of the available resources. The general workflow was separated into clear stages each of which had its goals, deliverables, and deadlines.

Project Schedule

The first step was the literature review where a gap in research was determined on the analysis of skin lesions, multimodal learning, and federated learning. The next step was the data collection, data pre-processing and data exploration analysis to verify quality of the data and its appropriateness to train the model.

In the next stage of development, the centralized deep learning models were introduced and tested as the default systems. Multimodal fusion strategies and segmentation elements were incorporated to the structure after developing baseline performance. The last step entailed designing and deploying the federated learning system and uncertainty-aware aggregation system. The training, evaluation, and performance analysis of the model were done in steps to optimize and refine the architecture. The project schedule was designed in a manner that the experimentation and analysis would be done way ahead of the final submission date.

Resource Management

The management of the resources also became critical to the project as computational requirements of deep learning models are high. Model development, debugging, and primary experimentation were done on high-performance local hardware with a dedicated GPU. In order to confirm scalability and reproducibility, chosen experiments were run on the cloud-based GPU platforms. Open-source libraries like PyTorch, torchvision, and timm were used as software resources and ensured flexibility, reliability, and maintainability. Scheduling of experiments and using previously trained models with care were used to keep the unnecessary computational overhead to a minimum.

Budget Management

The project was built keeping in mind the need to be cost effective through mainly using the hardware resources available and software tools that were available free of charge. There was no extra financial input in obtaining the dataset since all datasets were obtained on freely available and ethical repositories. Calculation was done sparingly using the cloud-based computational resources to prevent high costs. In total, the budget was also managed effectively through the balance of local and cloud resources, which made the project cost effective in an academic research context.

3.7 Risk Management

There are a number of technical, ethical, and functional risks associated with the development of an uncertain aware federated multimodal learning model to perform medical image analysis. These risks should be identified as early as possible, and corresponding mitigation measures should be included in the system to make sure that it is reliable, safe, and feasible.

The significant technical risk is that the performance of the model may deteriorate as a result of the heterogeneity of the data of various clients in the federated learning environment. Medical datasets obtained over various sources can also be different in terms of image quality, proportion of classes and the standard of annotation, which can adversely impact convergence and generalization. To reduce this risk, the proposed framework will use uncertainty-aware aggregation whereby the updates made by the clients will be weighted as per predictive confidence. This diminishes the effects of locally noisy and untrustworthy models and stabilizes global optimization.

The other risk area is the overfitting and biased predictions especially caused by class imbalance and underrepresentation of some skin types or rare types of lesions. This is tackled by means of class-weighted loss functions, macro-averaged evaluation measures and uncertainty estimation which discourage the overconfident prediction on classes that are under-represented and encourages a more balanced model behavior.

Medical AI research is prone to privacy and data security risks. Even though federated learning does not involve direct information exchange, there are still threats of model inversion or gradient leakage-associated information. To address this, the encrypted model updates are shared and no raw patient data are shared and stored at the center. Moreover, they perform experiments on ethically approved anonymized datasets only.

Transfer learning, early stopping, encoder freezing, and selective experimentation help mitigate operational risks that could affect the learning process, including excessive computational cost, long training properly, or hardware constraints. These policies cut down cost of resources and still performance stays. Lastly, the threat of misunderstanding in the medical practice is mitigated by explicitly stating the system as a decision-assistance tool and not a diagnostic substitution with uncertain and informed outputs that facilitate informed human judgment.

3.8 Economic Analysis

The research paper is meant to be cost effective in an academic research setting and the potential of the research paper is high when applied in practical health care settings. The economic analysis takes into account the cost of developing the proposed system and also the benefits of the cost in terms of long term.

In terms of development, the actual financial expense of the research is low. All data that is employed in this study is found in publicly accessible and ethical approved repositories, which removes the data acquisition costs. The software stack is fully based on open-source software including Python, PyTorch, torchvision, and timm which eliminates the cost of a license. The majority of the training and experimentation is done on the local hardware available, and little use of the cloud-based GPU resources is considered to be required to perform scalability and reproducibility verification to support control and cost-efficient use of computation.

Even though deep learning models based on transformers demand high-computational resources, efficiency-related design decisions, including transfer learning, pretrained encoders freezing, early stopping, and uncertainty-based aggregation, are implemented in the project. The methods save much training time, energy and cost of operation when compared to fully retrained centralized models.

Concerning the economic effects in the long term, the presented framework can help to decrease expenditures in healthcare by helping to detect severe skin-related issues, such as melanoma, earlier and more precisely. Early diagnosis will minimize the cost of treatment, decrease the number of complicated medical procedures and lessen the level of patients in the hospitals. Moreover, federated learning enables the institutions to collectively enhance diagnostic models without exchanging confidential data, which lowers the legal, administrative, and infrastructure expenses related to centralized data management.

In general, the cost-benefit ratio of this research is positive. The cheap development and the possibility of scalable deployment and healthcare cost reduction make the proposed system cost-efficient and appealing to the academic research and future clinical decision-support applications.

Chapter 4

Proposed Methodology

4.1 Methodology Overview

4.1.1 Research Workflow

This research proposes an *uncertainty-aware federated dual-branch multimodal learning framework* for privacy-preserving skin lesion diagnosis. The framework is designed to progressively address the limitations of centralized learning, single-task classification, and naive federated aggregation in heterogeneous medical environments. An overview of the complete pipeline is illustrated in Fig. 4.1, which highlights the interaction between model components and the federated optimization process.

The proposed framework is developed through four progressive stages, each introducing a new capability while building upon the previous one. These stages represent a conceptual research progression, ranging from centralized image-only learning to uncertainty-aware federated optimization.

For implementation and experimental consistency, the training process is organized into multiple execution phases, such as unimodal encoder training, centralized fusion learning, and federated optimization. These implementation phases correspond to, but do not redefine, the conceptual stages of the proposed framework.

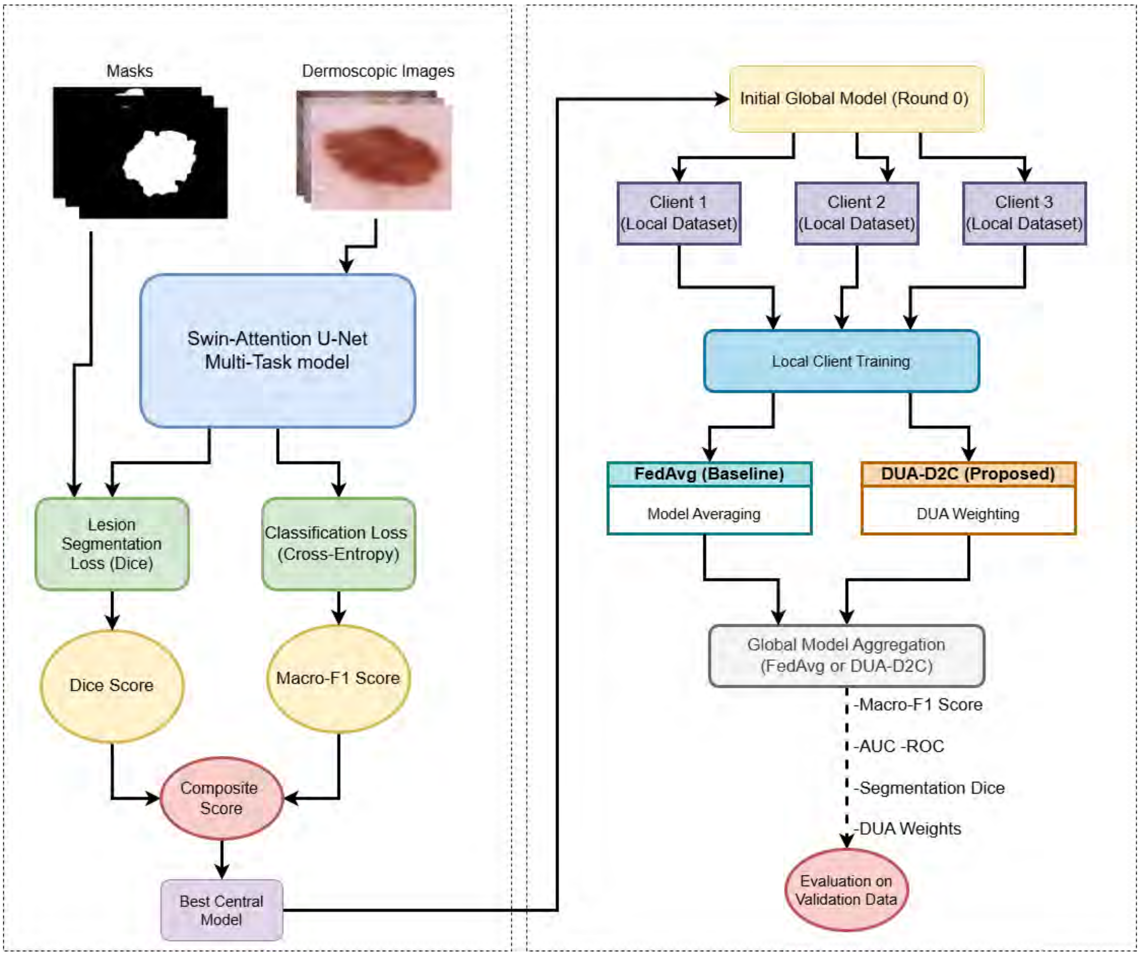


Figure 4.1: Overview of the proposed Swin-Attention U-Net based federated multi-task framework using FedAvg [32] DUA-D2C [49] aggregation.

4.1.2 Backbone Selection Rationale

The Swin Transformer [19] selected as the primary visual encoder in the proposed framework based on empirical performance observed during preliminary experimentation. Multiple state-of-the-art convolutional and transformer-based architectures, including EfficientNet variants and lightweight CNN backbones, were evaluated under identical training conditions. Among the evaluated models, Swin consistently achieved superior classification performance and demonstrated improved feature representation stability across datasets. Its hierarchical design and window-based self-attention mechanism further contribute to effective modeling of both local and global visual patterns, which is particularly beneficial for dermoscopic image analysis. Hence, Swin was adopted as the backbone architecture for all subsequent centralized and federated experiments to ensure a strong and consistent feature extraction foundation. A detailed quantitative comparison of candidate backbone architectures is provided in Chapter 5.

4.1.3 Stage I: Centralized Image-Only Classification

In the first stage, a centralized image-only classifier is trained using dermoscopic images. A Swin Transformer [19] is employed as the backbone network to extract hierarchical visual representations, owing to its strong performance in medical image analysis tasks. The extracted features are passed to a classification head to predict lesion categories.

This stage serves as a baseline reference, establishing the performance achievable using visual information alone without auxiliary supervision, multimodal fusion, or federated constraints. While effective, this approach lacks explicit lesion localization and does not account for predictive uncertainty or data privacy requirements.

4.1.4 Stage II: Centralized Dual-Branch Multi-Task Learning

To enhance spatial awareness and representation robustness, the second stage introduces a dual-branch multi-task learning architecture. The model consists of:

- A shared Swin Transformer [19] encoder that extracts high-level visual features from the input image;
- A classification branch that predicts lesion categories;
- A segmentation branch implemented using an Attention U-Net [12], which generates lesion masks to explicitly model lesion boundaries and salient regions.

The segmentation branch provides auxiliary supervision that guides the shared encoder to focus on diagnostically relevant regions, thereby improving classification accuracy and calibration. Both tasks are optimized jointly using a multi-task loss function, enabling shared feature learning while preserving task-specific objectives.

This stage establishes the dual-branch design central to the proposed framework and demonstrates the benefit of incorporating lesion localization as an auxiliary task.

4.1.5 Stage III: Federated Multi-Task Learning with FedAvg

In real-world clinical scenarios, medical data are distributed across multiple institutions and cannot be centrally aggregated due to privacy and regulatory constraints. To reflect this setting, the third stage extends the dual-branch multi-task model to a federated learning (FL) paradigm.

Each client (e.g., hospital or medical center) maintains a local copy of the multi-task model and performs training on its private dataset. Model updates are periodically transmitted to a central server, where they are aggregated using the Federated Averaging (FedAvg) [32] algorithm. The aggregated global model is then redistributed to all clients for subsequent training rounds.

Although FedAvg enables collaborative learning without direct data sharing, it assumes equal client reliability and is sensitive to non-IID data distributions, which are common in medical datasets. These limitations motivate the final stage of the proposed framework.

4.1.6 Stage IV: Uncertainty-Aware Federated Aggregation (DUA-D2C)

The final stage introduces **DUA-D2C** [49], an uncertainty-aware federated aggregation strategy designed to improve robustness under client heterogeneity. Instead of assigning uniform weights to all client updates, the proposed method evaluates each client model based on:

- **Predictive performance**, measured using the macro-averaged F1 score;
- **Predictive uncertainty**, quantified via entropy computed from the class probability distributions.

These two factors are combined into a unified client utility score, which determines the aggregation weight assigned to each client during global model updates. Clients exhibiting high predictive uncertainty or unreliable performance are automatically down-weighted, leading to a more stable and well-calibrated global model.
By integrating uncertainty estimation with federated optimization, DUA-D2C [49] adaptive, reliability-aware collaboration across distributed clients while preserving data privacy.

4.2 Dataset Overview

4.2.1 HAM10000

Primary Dataset

The ***HAM10000 (Human Against Machine)*** [14] dataset is one of the most comprehensive and widely used dermatoscopic image collections for pigmented skin lesion analysis. It contains 10,015 high-quality dermatoscopic images, collected over a period of 20 years from two different sites:

- The Department of Dermatology, Medical University of Vienna, Austria
- The Skin Cancer Practice of Cliff Rosendahl, University of Queensland, Australia

Images from the dataset were captured using multiple dermatoscopic systems that includes the Heine Dermaphot for analog diapositives ,the MoleMax HD digital dermatoscopy system, also the DermLite DL3 and lastly, Fluid models. These images were obtained under both polarized and non-polarized lighting conditions, using immersion fluids such as ethanol gel or ultrasound gel. All images were standardized to 800×600 pixels, and manual histogram correction was applied to enhance the color and the consistency of the contrast. The images show skin lesions from people of different backgrounds in Europe and Australia. They captured everything like in the real-world conditions such as body hair and skin texture. The dataset is publicly available on the ISIC Archive under a CC BY-NC 4.0 license for academic and non-commercial use. The dataset is has a high class-imbalanced as shown in Table 4.1 with melanocytic nevi making up about 67% of all samples, while other rare classes such as dermatofibroma and vascular lesions are about <2% each.

Class	Description	Number of Images
akiec	Actinic Keratoses / Intraepithelial Carcinoma	327
bcc	Basal Cell Carcinoma	514
bkl	Benign Keratosis-like Lesions	1,099
df	Dermatofibroma	115
mel	Melanoma	1,113
nv	Melanocytic Nevi	6,705
vasc	Vascular Lesions	142

Table 4.1: Class Distribution (HAM10000 [14] Dataset)

In this research, HAM10000 [14] serves as the primary dataset for evaluating the proposed dual-branch multi-task learning framework. The availability of segmentation masks allows the integration of an Attention U-Net [12] segmentation branch alongside the classification branch, facilitating explicit lesion localization and improved feature representation learning. The class distribution of the HAM10000 dataset, highlighting the presence of significant class imbalance across lesion categories, is illustrated in Figure 4.2.

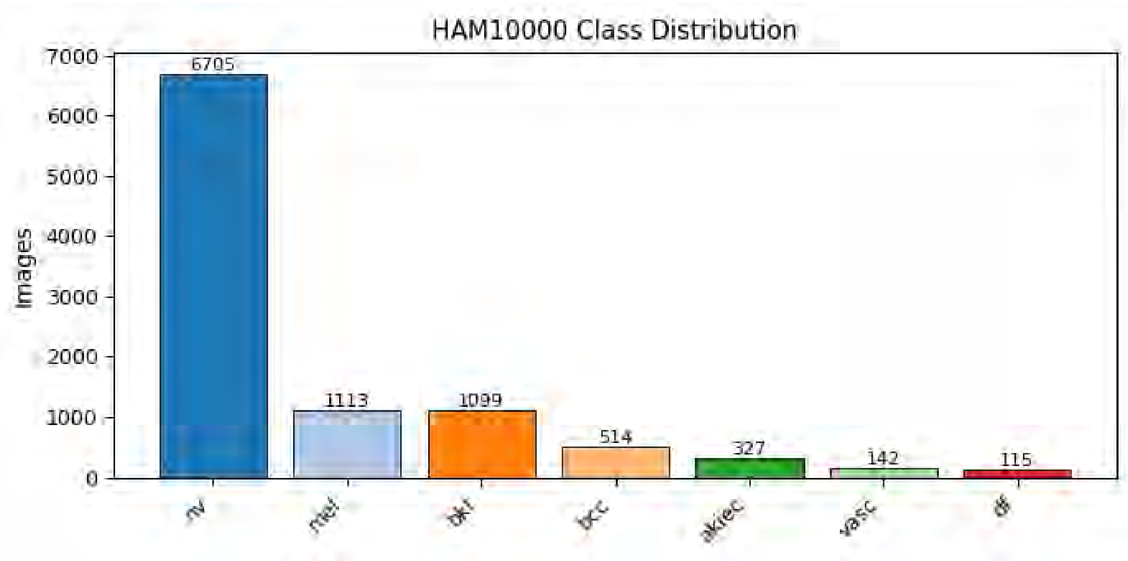


Figure 4.2: Class distribution of MILK10K Dataset

4.2.2 MILK10K

Secondary Dataset

The second dataset used in this study is the ***MILK10K*** [48] dataset. The dataset contains 10,480 images with each case includes a clinical close-up and a matching dermatoscopic image of 5,240 lesions. It is mainly used for multi-class lesion classification tasks in deep learning research, and not for segmentation. The dataset was curated from multiple public repositories such as:

- ISIC Archive
- Derm7pt
- PH2
- PAD-UFES-20 datasets

All the images were standardized in terms of aspect ratio, orientation, and resolution. Dermatoscopic images were taken using polarized light, and close-up photos were captured with various professional and clinical cameras.

The metadata of this dataset has patient age, sex, skin tone, lesion site, and diagnosis, along with the method of confirmation (mostly histopathology, used in 95.7% of cases). Skin tone is divided into six levels from very dark (0) to very light (5).

The dataset includes 11 broad diagnostic categories. Also, the MILK10K dataset is heavily imbalanced as shown in Table 4.2 with Basal Cell Carcinoma (BCC) forming nearly half of all samples, followed by melanocytic nevi (NV) and benign keratinocytic lesions (BKL).

Class	**Description**	**Number of Lesions**
BCC	Basal Cell Carcinoma	2,522
NV	Melanocytic Nevus (all types)	746
BKL	Benign Keratinocytic Lesions	544
SCC/KA	Squamous Cell Carcinoma / Keratoacanthoma	473
MEL	Melanoma	450
AK/IEC	Actinic Keratosis / Intraepidermal Carcinoma	303
DF	Dermatofibroma	52
INF	Inflammatory & Infectious Lesions	50
VASC	Vascular Lesions & Hemorrhage	47
BEN_OTH	Other Benign Proliferations (including collisions)	44
MAL_OTH	Other Malignant Proliferations (including collisions)	9

Table 4.2: Lesion Classes Distribution (Milk10K [48] Dataset)

In this research, MILK10k [48] is used to evaluate the generalization capability of the proposed framework under multimodal and federated learning scenarios without segmentation supervision. Accordingly, experiments conducted on MILK10k employ classification-oriented model variants, omitting the segmentation branch while retaining the same backbone architecture for consistency. The class distribution of the MILK10k dataset, illustrating the inherent class imbalance across diagnostic categories, is shown in Figure 4.3.

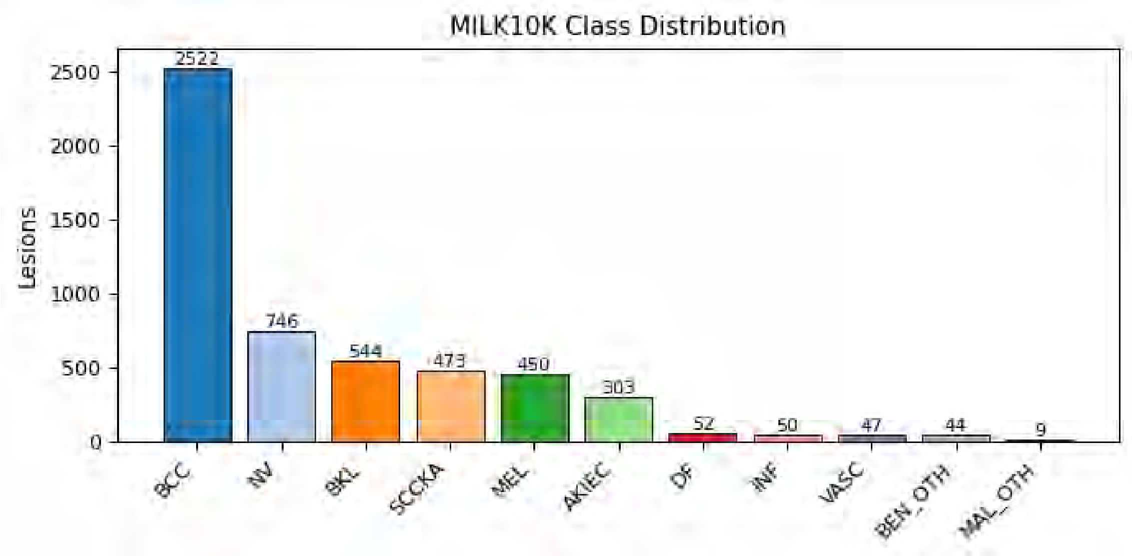


Figure 4.3: Class distribution of MILK10K Dataset

Lastly, the Table 4.3 shows a brief comparison summary of both the datasets

Feature	**HAM10000**	**MILK10K**
Total Images	10,015	∼10,000
Number of Classes	7	10
Image Type	Dermatoscopic	Dermatoscopic + Clinical
Resolution	800×600 px	224–256×256 px (standardized)
Segmentation Masks	Available	Not available
Collected From	Vienna (Austria), Queensland (Australia)	Multiple datasets (ISIC, Derm7pt, PH2, PAD-UFES-20)
Primary Tasks	Segmentation + Classification	Classification only
Verification	Histopathology, Confocal Microscopy, Expert Consensus	Histopathology, Expert Annotation, Dataset Integration

Table 4.3: Summary comparison between HAM10000 [14] and MILK10K [48] dataset

4.2.3 Data Preprocessing and Augmentation

Before training and evaluation, all the images in the datasets underwent a standardized preprocessing pipeline to ensure consistency across modalities. The processes performed were metadata extraction, image normalization and resizing. Two datasets were primarily used: MILK10k [48] for multimodal classification and HAM10000 [14] segmentation and classification. Even though the preprocessing logic remained largely consistent, but each dataset has different structure and file formats ,so it required specific handling.

HAM10000 Dataset Preprocessing

The HAM10000 [14] dataset in this study was used for lesion segmentation and classification. Each lesion image in this dataset has a corresponding ground-truth segmentation mask

1. **Image-Mask Pairing**

   All available image files were scanned, and corresponding mask files were identified. Then only the pairs for which both image and mask existed were included in the final dataset. Any image without a corresponding mask (or vice versa) was excluded to maintain one-to-one correspondence between input and ground truth.

2. **Image Resizing**

   Both the lesion images and their corresponding segmentation masks were resized **224 × 224 pixels**. This ensured uniform spatial dimensions and computational efficiency during training.

3. **Normalization**

   The image tensors were normalized using ImageNet mean and standard deviation values::

   $$\text{mean} = [0.485, 0.456, 0.406], \quad \text{std} = [0.229, 0.224, 0.225]$$

   The segmentation masks, on the other hand, were not normalized since they contained binary values representing lesion presence (1) or background (0).

MILK10K Dataset Preprocessing

For each lesion in the MILK10K [48] dataset there are two image modalities — one *clinical* and one *dermoscopic*. Each sample has a one-hot encoded label that represents one of 11 diagnostic classes:

```
["AKIEC", "BCC", "BEN_OTH", "BKL", "DF", "INF",
   "MAL_OTH", "MEL", "NV", "SCCKA", "VASC"]
```

Two metadata files were first loaded:

- `MILK10k_Training_Metadata.csv` — This contains lesion and image identifiers, along with the corresponding image modality (*clinical* or *dermoscopic*).
- `MILK10k_Training_GroundTruth.csv` — This contains the one-hot encoded ground-truth labels for each lesion.

Before training, the pipeline validated that both images (clinical and dermoscopic) existed for each lesion and were not corrupted. For each lesion, a unified data structure was created by merging information from both metadata sources. This structure links the *clinical* and *dermoscopic* images with their corresponding diagnostic label, ensuring a consistent one-to-one correspondence between modalities throughout data loading, augmentation, and training.

So during data preparation, We made sure:

- The same lesion ID is used for both image types.
- The clinical path and dermatology path are stored together under that ID.
- The label (diagnosis) is attached once — and applies to both images.

This structure ensures that both clinical and dermoscopic images always remain synchronized throughout data loading, augmentation, and training, maintaining the one-to-one correspondence between modalities.

1. **Image Resizing**

   Both clinical and dermoscopic images were resized to **224 × 224 pixels**, aligning with the input dimensionality requirements of convolutional neural networks such as ResNet [3] and VGG [5].

2. **Normalization**

   To maintain compatibility with pretrained ImageNet models, all images were normalized using the ImageNet channel-wise mean and standard deviation:

   $$\text{mean} = [0.485, 0.456, 0.406], \quad \text{std} = [0.229, 0.224, 0.225]$$

   This transformation standardized the input distribution which enhanced transfer learning performance.

HAM10000 Data Augmentation

Data augmentation for the HAM10000[14] dataset was designed differently for the segmentation and classification tasks. An overview illustration is shows in figure 4.4

1. **Segmentation Task**

   For segmentation, augmentation was deliberately minimal to maintain **geometric consistency** between the lesion image and its corresponding binary mask. Excessive augmentation could cause spatial misalignment, which would impact the mask accuracy negatively. The applied transformations included:

   - **RandomHorizontalFlip / RandomVerticalFlip :** Applied sparingly to increase spatial diversity while maintaining mask alignment.

   - **RandomRotation :** Small rotational variations (up to $\pm 15°$) were occasionally applied to simulate orientation changes.

   No color-based transformations (e.g., brightness or contrast adjustments) were used, as they could introduce inconsistencies between the image and mask pair. The segmentation masks **were not normalized** so that they retain their binary values (0 and 1) but were resized to match the image dimensions.

2. **Classification Task**

   For the classification task based on the **HAM10000 Task 3** [14] dataset, a more extensive augmentation strategy was employed to increase model robustness and reduce overfitting. The following transformations were applied:

   - **RandomHorizontalFlip:** Introduced left-right variability.

   - **RandomRotation ($\pm 15°$):** Simulated different lesion orientations.

   - **ColorJitter:** Randomly altered brightness and contrast to increase color variation tolerance.

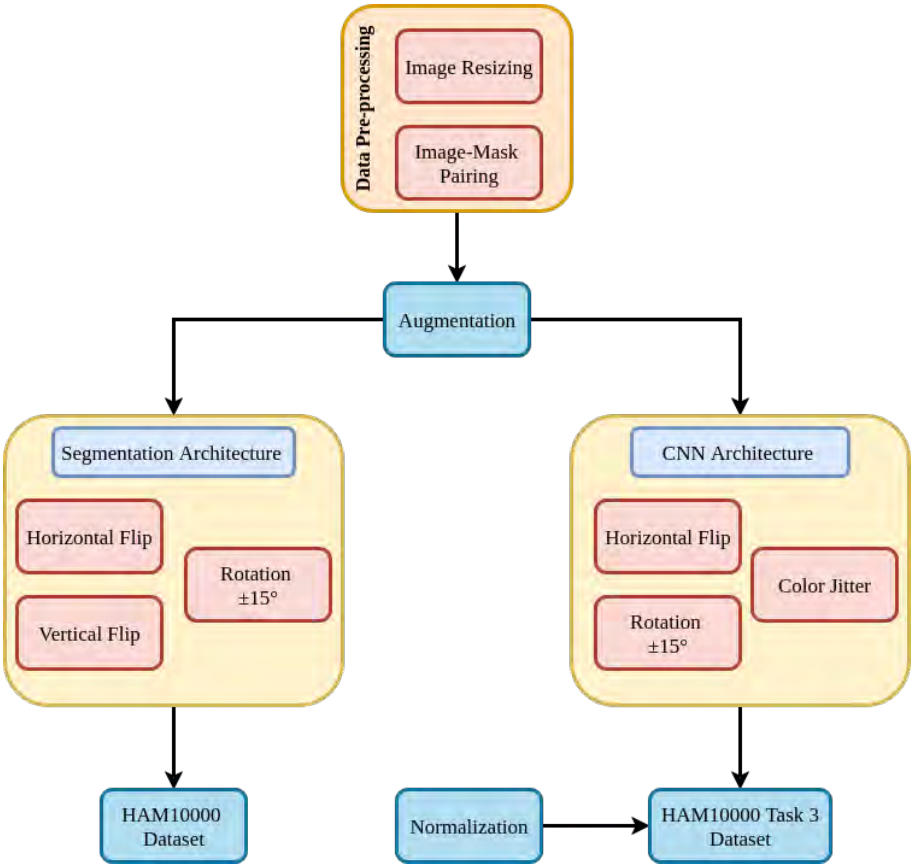


Figure 4.4: Step-by-step data augmentation process with normalization applied

4.2.4 Training Configurations

The training configurations presented in Tables 4.4–4.7 correspond to the four progressive stages of the proposed research framework and are designed to systematically evaluate the contribution of each modeling component.

Table 4.4 the configuration for Stage I, which establishes a centralized image-only baseline using a Swin Transformer [19] backbone. This setting serves as a lower-bound reference, quantifying the performance achievable from visual information alone without auxiliary supervision or federated constraints.

Parameter	Setting
Dataset	HAM10000 [14], MILK10k [48]
Input modality	Dermoscopic images only
Backbone	Swin Transformer [19] (ImageNet pretrained)
Task	Multi-class skin lesion classification
Number of classes	7 (HAM10000)
Loss function	BCEWithLogitsLoss (one-hot targets)
Optimizer	AdamW
Initial learning rate	3×10^{-4}
Batch size	16
Maximum epochs	30
Early stopping	Patience = 5 epochs
Validation metric	Macro-F1 score

Table 4.4: Stage I configuration: Centralized image-only Swin-based classification.

Table 4.5 details the configuration for Stage II, where a dual-branch multi-task architecture is introduced by coupling classification with lesion segmentation using an Attention U-Net [12]. This stage evaluates the impact of explicit lesion localization as auxiliary supervision and forms the foundation for all subsequent federated experiments. Due to the requirement for pixel-level annotations, this configuration is applied exclusively to the HAM10000[14] dataset.

Parameter	Setting
Dataset	HAM10000
Input modality	Dermoscopic images
Shared encoder	Swin Transformer [19]
Classification branch	Linear classifier (7 classes)
Segmentation branch	Attention U-Net [12]
Classification loss ($\mathcal{L}_{cls}$)	BCEWithLogitsLoss
Segmentation loss ($\mathcal{L}_{seg}$)	Dice loss
Multi-task objective	$\mathcal{L} = \mathcal{L}_{cls} + \alpha\mathcal{L}_{seg}$
Optimizer	AdamW
Batch size	16
Maximum epochs	30
Early stopping	Patience = 5 epochs

Table 4.5: Stage II configuration: Centralized dual-branch multi-task learning using Swin [19] and Attention U-Net [12].

The federated learning setup is introduced in Stage III, with its configuration summarized in Table 4.6. Here, the dual-branch model is deployed across multiple simulated clients and trained collaboratively using the Federated Averaging (FedAvg) [32] algorithm. This stage reflects a privacy-preserving learning scenario commonly encountered in clinical practice and serves as the federated baseline against which more advanced aggregation strategies are compared.

Parameter	**Setting**
Dataset	HAM10000
Federated setting	Cross-silo (simulated hospitals)
Number of clients	3
Local model	Dual-branch Swin [19] + Attention U-Net [12]
Local epochs per round	1
Global communication rounds	15
Local optimizer	AdamW
Local learning rate	1×10^{-3}
Aggregation method	Federated Averaging (FedAvg) [32]
Client weighting	Uniform
Shared parameters	Model head parameters
Validation metric	Macro-F1 score

Table 4.6: Stage III configuration: Federated multi-task learning with FedAvg [32] aggregation.

Finally, Table 4.7 presents the configuration for Stage IV, which incorporates the proposed DUA-D2C [49] uncertainty-aware aggregation strategy. In this setting, client updates are weighted based on a combination of predictive performance and uncertainty, enabling adaptive and reliability-aware model aggregation under heterogeneous data distributions. This stage represents the full realization of the proposed framework and directly addresses the challenges of non-IID data and client variability in federated medical learning.

Parameter	Setting
Base federated method	FedAvg [32]
Uncertainty estimation	Predictive entropy (softmax probabilities)
Performance metric (a_i)	Macro-F1 score
Uncertainty metric (u_i)	Normalized inverse entropy
Client utility function	$s_i = \lambda a_i + (1 - \lambda) u_i$
Trade-off parameter (λ)	0.7
Aggregation rule	Weighted model averaging
Weight normalization	$\sum_i w_i = 1$
Stability safeguard	Uniform FedAvg fallback
Evaluation focus	Robustness under non-IID clients

Table 4.7: Stage IV configuration: Uncertainty-aware federated aggregation using DUA-D2C.

4.3 Model Design Specification

4.3.1 Overview

This section presents the overall design of the proposed model architecture, which constitutes a unified, uncertainty-aware federated dual-branch learning framework for skin lesion diagnosis. The architecture is designed to be modular and extensible, allowing different task-level and modality-level components to be selectively activated depending on dataset characteristics and clinical requirements.

At its core, the framework employs a shared Swin Transformer encoder [19] to extract hierarchical visual representations from dermoscopic or clinical images. Building upon this shared representation, two complementary forms of dual-branch learning are incorporated. The first is a task-level dual-branch design, where the encoder output is simultaneously optimized for lesion classification and lesion segmentation through a classification head and an Attention U-Net [12] decoder, respectively. This multi-task formulation encourages spatially informed feature learning by explicitly modeling lesion boundaries and salient regions.

The second form is a modality-level dual-branch design, where visual features extracted by the Swin encoder are fused with structured patient metadata, such as demographic and clinical attributes. A lightweight metadata encoder maps non-image inputs into a latent feature space, which is then integrated with image features through configurable fusion mechanisms. This multimodal fusion enables the model to leverage complementary clinical information alongside visual cues.

To support privacy-preserving collaborative learning across distributed medical institutions, the proposed architecture is deployed within a federated learning paradigm.

Each client maintains a local instance of the model and performs training on private data, while a central server coordinates model aggregation. Unlike conventional federated averaging, the proposed framework incorporates an uncertainty-aware aggregation strategy, termed DUA-D2C [49], which adaptively weights client contributions based on predictive performance and uncertainty estimates. This design enhances robustness under non-identically distributed data and mitigates the influence of unreliable or overconfident client updates.

Importantly, the architecture is formulated as a superset model: when segmentation annotations are unavailable, the segmentation branch can be disabled; when metadata is absent, the modality fusion component is omitted. This flexibility allows the same architectural blueprint to be consistently applied across multiple datasets, including HAM10000 [14] and MILK10k [48] , while preserving methodological coherence.

An overview of the complete architecture, illustrating the interaction between task-level branching, modality-level fusion, and uncertainty-aware federated aggregation, is depicted in Fig.4.5

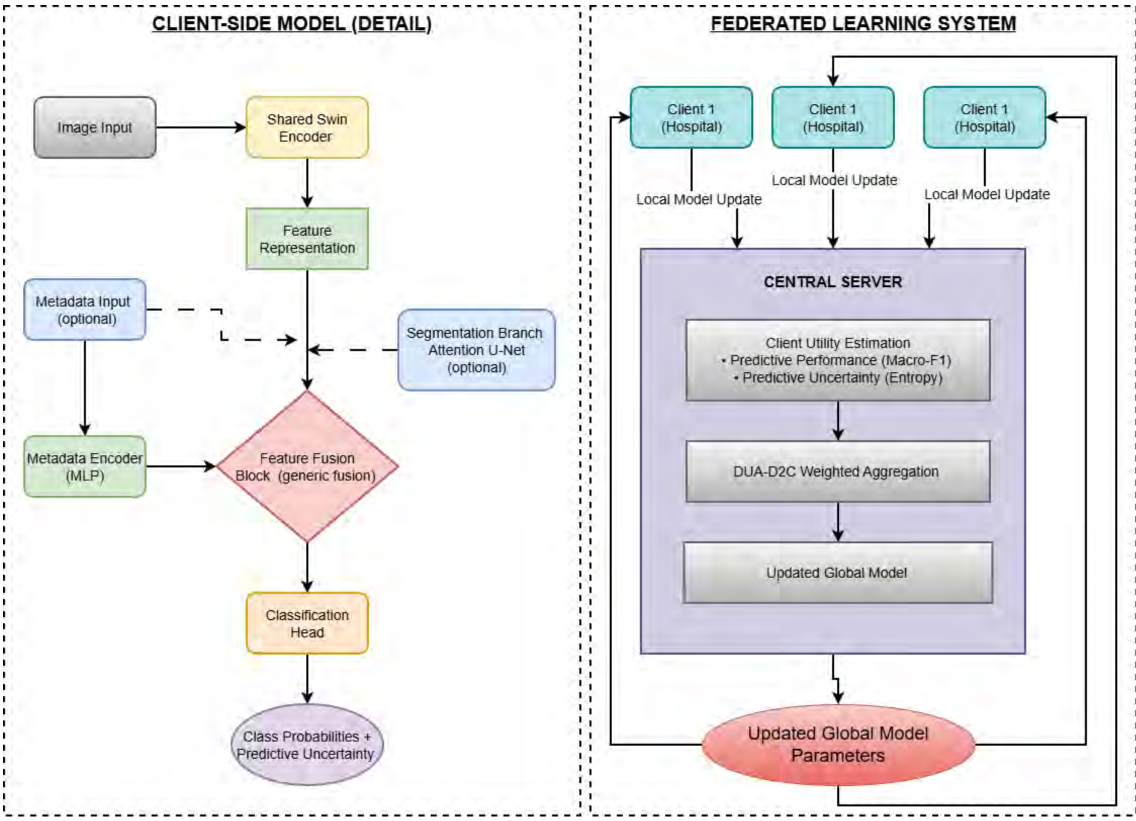


Figure 4.5: Conceptual overview of the proposed uncertainty-aware federated dual-branch architecture, illustrating task-level multi-task learning, modality-level feature fusion, and DUA-D2C-based federated aggregation.

4.3.2 Architecture Description

This section presents the detailed architecture of the proposed uncertainty-aware federated dual-branch multimodal framework. The architecture is designed to flexibly support centralized and federated learning, single-task and multi-task supervision, and unimodal as well as multimodal data configurations. Its modular design enables selective activation of components depending on dataset characteristics, while preserving a unified learning formulation.

Shared Visual Encoder: Swin Transformer

At the core of the proposed framework lies a Swin Transformer [19], which serves as the shared visual encoder for all experiments.
Given an input image

$$\mathbf{x} \in \mathbb{R}^{H \times W \times 3}, \tag{4.1}$$

the image is first partitioned into non-overlapping patches of size $P \times P$. Each patch is flattened and projected into an embedding space, forming a sequence of patch embeddings:

$$\mathbf{z}_0 = [\mathbf{x}_p^1, \mathbf{x}_p^2, \dots, \mathbf{x}_p^N]\mathbf{E}, \tag{4.2}$$

where $N = HW/P^2$ and $\mathbf{E}$ is a learnable linear projection matrix.
The Swin Transformer processes these embeddings through multiple hierarchical stages, progressively reducing spatial resolution while increasing feature dimensionality. Within each stage, attention is computed locally within fixed-size windows using window-based multi-head self-attention (W-MSA):

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{B}\right)\mathbf{V}, \tag{4.3}$$

where $\mathbf{B}$ denotes relative positional bias and d is the attention head dimension.
To enable cross-window interactions, successive layers alternate between regular and shifted window configurations. Each Swin Transformer block follows a residual structure:

$$\hat{\mathbf{z}}^{(l)} = \text{W-MSA}(\mathrm{LN}(\mathbf{z}^{(l-1)})) + \mathbf{z}^{(l-1)}, \tag{4.4}$$

$$\mathbf{z}^{(l)} = \mathrm{MLP}(\mathrm{LN}(\hat{\mathbf{z}}^{(l)})) + \hat{\mathbf{z}}^{(l)}, \tag{4.5}$$

where LN denotes layer normalization and the MLP consists of two fully connected layers with GELU activation.
After the final Swin stage, global average pooling is applied to obtain a compact visual representation:

$$\mathbf{f}_v = \mathrm{GAP}(\mathbf{z}_L) \in \mathbb{R}^D. \tag{4.6}$$

This feature vector forms the shared latent representation for downstream tasks.

Task-Level Dual-Branch Learning

To enhance representation robustness and spatial awareness, the architecture supports task-level dual-branch learning, enabling joint optimization of classification and segmentation objectives.

Classification Branch The classification branch operates on the shared feature representation $\mathbf{f}$ (which may correspond to image-only or fused multimodal features). Class logits are computed as:

$$\mathbf{o} = \mathbf{W}_c\mathbf{f} + \mathbf{b}_c, \tag{4.7}$$

where $\mathbf{W}_c$ and $\mathbf{b}_c$ are learnable parameters. Final class probabilities are obtained using a sigmoid activation:

$$\hat{\mathbf{y}} = \sigma(\mathbf{o}), \tag{4.8}$$

allowing independent prediction of multiple diagnostic categories.

Segmentation Branch Using Attention U-Net (Optional) When pixel-level annotations are available, an auxiliary segmentation branch is activated using an Attention U-Net [12] decoder. Intermediate feature maps from multiple Swin stages are forwarded to the decoder through skip connections, enabling multi-scale feature fusion.
Attention gates suppress irrelevant background activations and emphasize lesion-relevant regions. The gated feature map at scale i is computed as:

$$\mathbf{G}_i = \alpha_i(\mathbf{F}_i, \mathbf{g}) \odot \mathbf{F}_i, \tag{4.9}$$

where $\mathbf{F}_i$ denotes the encoder feature map at scale i, $\mathbf{g}$ is a gating signal from deeper layers, and $\alpha_i(\cdot)$ denotes the learned attention coefficients. The decoder produces a dense segmentation mask via convolutional upsampling and softmax activation.
This auxiliary task acts as a regularizer for the shared encoder by enforcing spatial consistency and lesion localization. When segmentation labels are unavailable, this branch is disabled without modifying the remaining architecture.

Modality-Level Dual-Branch Learning

In addition to task-level branching, the proposed framework incorporates modality-level dual-branch learning to integrate visual information with structured patient metadata.

Metadata Encoder Structured metadata such as age, sex, lesion location, and clinical attributes are represented as a vector

$$\mathbf{m} \in \mathbb{R}^M. \tag{4.10}$$

A lightweight multi-layer perceptron encodes this information into a latent representation:

$$\mathbf{f}_m = \phi(\mathbf{W}_m\mathbf{m} + \mathbf{b}_m), \tag{4.11}$$

where $\phi(\cdot)$ denotes a nonlinear activation function.

Feature Fusion The encoded visual and metadata features are combined using a generic fusion operation:

$$\mathbf{f}_{\text{fusion}} = \mathcal{F}(\mathbf{f}_v, \mathbf{f}_m), \tag{4.12}$$

where $\mathcal{F}(\cdot)$ may correspond to simple concatenation, gated fusion, or feature-wise linear modulation (FiLM). This modular design enables systematic evaluation of different fusion strategies without altering the core architecture.

Predictive Uncertainty Estimation

To quantify predictive uncertainty, the framework computes entropy over the predicted class probabilities:

$$\mathcal{H}(\hat{\mathbf{y}}) = -\sum_{k=1}^{K} \hat{y}_k \log(\hat{y}_k + \epsilon), \tag{4.13}$$

where ϵ is a small constant for numerical stability. Higher entropy values indicate increased uncertainty and reduced prediction confidence. These uncertainty estimates are later used for federated aggregation.

Federated Deployment and DUA-D2C Aggregation

The complete architecture is deployed in a federated learning setting, where multiple clients independently train local copies of the model on private data. At each communication round t, client i produces an updated parameter set $\theta_i^{(t)}$, along with a performance score a_i and an uncertainty score c_i.

The proposed DUA-D2C [49] aggregation strategy computes a unified client utility score:

$$s_i = \lambda a_i + (1 - \lambda)\tilde{c}_i, \tag{4.14}$$

where $\tilde{c}_i$ denotes normalized inverse entropy and λ controls the trade-off between accuracy and uncertainty. The global model is updated as:

$$\theta^{(t+1)} = \sum_{i=1}^{N} \frac{s_i}{\sum_{j=1}^{N} s_j} \theta_i^{(t)}. \tag{4.15}$$

By down-weighting unreliable or uncertain client updates, DUA-D2C improves robustness under heterogeneous data distributions while preserving data privacy.

Architectural Summary

The proposed architecture unifies hierarchical transformer-based visual encoding, task-level and modality-level dual-branch learning, multimodal fusion, and uncertainty-aware federated optimization into a single coherent framework. Its modular design allows selective activation of components based on dataset availability while maintaining methodological consistency across all experimental settings.

HAM10000 Architecture Instantiation

For the HAM10000 [14] dataset, the proposed framework is instantiated using a single visual input stream, as only dermoscopic images are available. As illustrated in Fig. 4.6, a Swin Transformer serves as the shared visual encoder to extract high-level image representations, while structured patient metadata (age, sex, and lesion localization) are encoded through a lightweight metadata encoder.

The resulting visual and metadata features are fused using a trainable fusion head, following the general multimodal formulation described in Section 4.3.2. The segmentation branch is inactive for this configuration due to the absence of pixel-level annotations. The fusion head parameters are optimized in a federated learning setting, where client updates are aggregated using the proposed DUA-D2C strategy to account for performance variability and predictive uncertainty across clients.

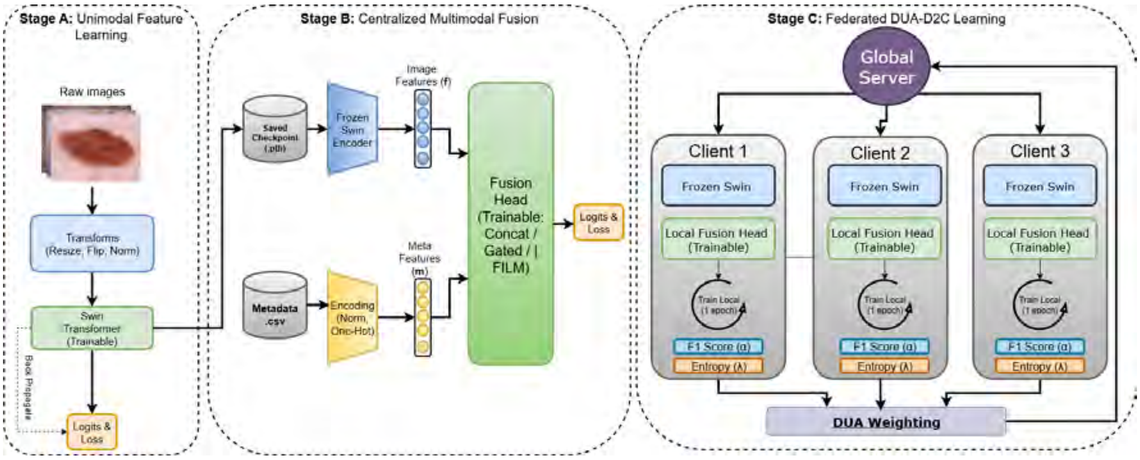


Figure 4.6: Architecture instantiation of the proposed framework for the HAM10000 [14] dataset, showing dermoscopic image encoding, metadata fusion, and federated optimization using the proposed DUA-D2C [49] strategy.

MILK10k Architecture Instantiation

The MILK10k [48] enables a richer multimodal configuration, incorporating both clinical and dermoscopic image modalities. As shown in Fig. 4.7, the proposed framework is instantiated with two parallel Swin Transformer [19] encoders, each dedicated to one visual modality. The modality-specific feature embeddings are combined with encoded patient metadata through a shared fusion head.

This configuration explicitly realizes modality-level dual-branch learning within the general architecture. Only the fusion head is trained during centralized and federated optimization, while the visual encoders remain frozen. Federated training is performed using the proposed DUA-D2C [49] strategy, allowing adaptive weighting of client contributions based on predictive performance and uncertainty.

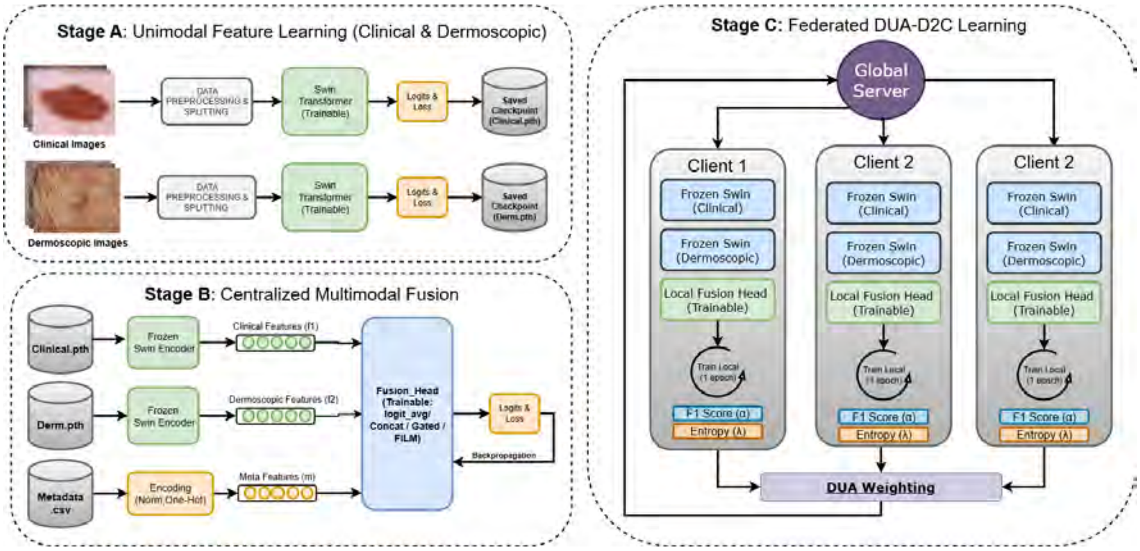


Figure 4.7: Architecture instantiation of the proposed framework for the MILK10k [48] dataset, illustrating dual visual encoders, metadata fusion, and federated learning with DUA-D2C [49].

4.3.3 Loss Functions and Optimization

This section describes the objective functions used for training the proposed framework and the optimization procedure under centralized and federated settings. The overall learning objective integrates (i) lesion classification, (ii) optional lesion segmentation (when pixel-level masks are available), and (iii) uncertainty-aware aggregation within the federated optimization loop.

Classification Loss

Let the model output classification logits be denoted by

$$\mathbf{o} = [o_1, o_2, \ldots, o_K] \in \mathbb{R}^K, \tag{4.16}$$

where K is the number of diagnostic categories, and let the corresponding ground-truth label vector be

$$\mathbf{y} = [y_1, y_2, \ldots, y_K] \in \{0, 1\}^K. \tag{4.17}$$

To support multi-label prediction (and to remain applicable when labels are represented as one-hot vectors), we employ the Binary Cross-Entropy with Logits loss (BCEWithLogitsLoss):

$$\mathcal{L}_{\text{cls}}(\mathbf{o}, \mathbf{y}) = -\frac{1}{K} \sum_{k=1}^{K} \Big(y_k \log \sigma(o_k) + (1 - y_k) \log \big(1 - \sigma(o_k)\big) \Big), \tag{4.18}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function.

Remark (single-label formulation). For datasets treated strictly as single-label multi-class classification, the logits $\mathbf{o}$ can alternatively be optimized with categorical cross-entropy using a softmax activation. In this work, BCEWithLogitsLoss provides a unified formulation compatible with one-hot encoded targets and multi-label scenarios.

Segmentation Loss (Optional)

When segmentation masks are available, the model produces a dense prediction map $\hat{\mathbf{S}} \in [0,1]^{H\times W}$ (binary mask) and is trained using a segmentation objective. Let $\mathbf{S} \in \{0,1\}^{H\times W}$ denote the ground-truth mask. We adopt a Dice loss to encourage overlap between predicted and ground-truth lesion regions:

$$\mathcal{L}_{\text{dice}}(\hat{\mathbf{S}}, \mathbf{S}) = 1 - \frac{2\sum_p \hat{S}_p S_p + \epsilon}{\sum_p \hat{S}_p + \sum_p S_p + \epsilon}, \tag{4.19}$$

where p indexes pixels and ϵ is a small constant for numerical stability.
Optionally, the segmentation loss can be augmented with pixel-wise binary cross-entropy:

$$\mathcal{L}_{\text{bce\_seg}}(\hat{\mathbf{S}}, \mathbf{S}) = -\frac{1}{HW}\sum_p \Big(S_p \log(\hat{S}_p) + (1-S_p)\log(1-\hat{S}_p)\Big), \tag{4.20}$$

yielding a combined segmentation objective:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{dice}} + \mathcal{L}_{\text{bce\_seg}}. \tag{4.21}$$

Multi-Task Objective

In the task-level dual-branch setting (classification + segmentation), the overall loss is defined as a weighted sum:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \alpha\, \mathcal{L}_{\text{seg}}, \tag{4.22}$$

where $\alpha \geq 0$ controls the influence of segmentation supervision. When segmentation annotations are unavailable, the segmentation branch is disabled and $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}}$.

Optimization in Centralized Training

In centralized training, model parameters θ are optimized by minimizing the empirical risk over the training set $\mathcal{D}$:

$$\theta^* = \arg\min_{\theta} \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x},\mathbf{m},\mathbf{y},\mathbf{S})\in\mathcal{D}} \mathcal{L}_{\text{total}}(\theta; \mathbf{x}, \mathbf{m}, \mathbf{y}, \mathbf{S}), \tag{4.23}$$

where $\mathbf{x}$ is the image input, $\mathbf{m}$ denotes optional metadata, $\mathbf{y}$ denotes the classification target, and $\mathbf{S}$ denotes the optional segmentation mask. Optimization is performed using AdamW with weight decay to stabilize training and improve generalization.

Federated Optimization with FedAvg

In federated learning, the training data are distributed across N clients, where client i holds a private dataset $\mathcal{D}_i$ and local model parameters θ_i. At global round t, the server broadcasts the current global parameters $\theta^{(t)}$ to all clients. Each client performs E_{local} epochs of local optimization and returns updated parameters $\theta_i^{(t)}$ to the server.

Under Federated Averaging (FedAvg [32]), the global model is updated as:

$$\theta^{(t+1)} = \sum_{i=1}^{N} \frac{|\mathcal{D}_i|}{\sum_{j=1}^{N} |\mathcal{D}_j|} \, \theta_i^{(t)}. \tag{4.24}$$

While FedAvg enables privacy-preserving collaboration, it assigns aggregation weights based primarily on dataset size and may be sensitive to client heterogeneity and non-IID data.

Uncertainty Quantification via Predictive Entropy

To quantify predictive uncertainty, we compute entropy from the predicted class probabilities. Let $\hat{\mathbf{p}} \in [0,1]^K$ denote the predicted probability vector. For multi-class probabilities, entropy is:

$$\mathcal{H}(\hat{\mathbf{p}}) = -\sum_{k=1}^{K} \hat{p}_k \log(\hat{p}_k + \epsilon), \tag{4.25}$$

where ϵ avoids numerical instability. Higher entropy indicates higher uncertainty. In practice, entropy is computed per sample and averaged over a validation set to obtain a client-level uncertainty measure.

Uncertainty-Aware Federated Aggregation (DUA-D2C)

The proposed DUA-D2C [49] assigns adaptive aggregation weights based on (i) predictive performance and (ii) predictive uncertainty. After local training, each client i is evaluated on a shared validation set (or a server-held reference set) to obtain:

- a performance score a_i (macro-averaged F1 score), and
- an uncertainty score c_i (mean predictive entropy).

To map uncertainty into a reliability signal, we compute an inverse-entropy quantity:

$$u_i = \frac{1}{c_i + \delta}, \tag{4.26}$$

where $\delta > 0$ prevents division by zero. The reliability scores are then normalized to $[0,1]$:

$$\tilde{u}_i = \frac{u_i - u_{\min}}{u_{\max} - u_{\min} + \kappa}, \tag{4.27}$$

where $\kappa > 0$ is a small constant for numerical stability.
A unified client utility score combines performance and uncertainty:

$$s_i = \lambda a_i + (1-\lambda)\tilde{u}_i, \tag{4.28}$$

with $\lambda \in [0,1]$ controlling the trade-off. The final aggregation weights are:

$$w_i = \frac{s_i}{\sum_{j=1}^{N} s_j + \varepsilon}, \tag{4.29}$$

where ε prevents degenerate normalization. The global model update becomes:

$$\theta^{(t+1)} = \sum_{i=1}^{N} w_i \, \theta_i^{(t)}. \tag{4.30}$$

This uncertainty-aware aggregation down-weights client updates that are either low-performing or highly uncertain, improving robustness under heterogeneous client distributions while preserving privacy by keeping data local.

4.3.4 Efficiency Analysis

The proposed framework is designed for efficient deployment in resource-constrained federated environments. Most parameters reside in the Swin Transformer backbone, which is pretrained centrally and kept frozen during both centralized and federated training. Consequently, only lightweight components—including the metadata encoder, fusion head, and optional segmentation decoder—are trained and communicated across clients, significantly reducing memory and communication overhead.

From a computational perspective, Swin Transformer employs window-based self-attention, enabling efficient feature extraction with reduced complexity compared to global self-attention. In multimodal settings (e.g., MILK10k), parallel visual encoders increase inference cost but do not affect federated training complexity, as these encoders remain frozen. The segmentation branch introduces additional computation only when pixel-level annotations are available.

In the federated setting, communication efficiency is achieved by transmitting updates for only the trainable components. The proposed DUA-D2C aggregation strategy incurs negligible overhead, relying solely on scalar performance and uncertainty statistics in addition to standard model updates.

4.4 Implementation of the Proposed DUA-D2C Framework

Algorithm 1 details the proposed DUA-D2C [49] strategy, which adaptively weights client updates using both predictive performance and uncertainty estimates.

Algorithm 1 Dynamic Uncertainty-Aware Divide2Conquer (DUA-D2C)

Input: Training data D, Validation data D_{val}; Num. subsets N; Epochs E, E_{global}; Batch B; Learning rate lr; Parameters: λ, C_{max}, δ_c, K, ϵ_s
Output: Central model M_c with aggregated weights θ_c

function DUA-D2C(D, D_{val}, N, E, E_{global}, B, lr, λ, C_{max}, δ_c, K, ϵ_s)
  Shuffle D; Divide D into N subsets: $\{D_1, \ldots, D_N\}$ maintaining class distributions

  Initialize central model M_c with weights θ_c
  Set min. inverse entropy: $inv\_\epsilon_{min} = 1/(\log(K) + \delta_c)$
  for $g = 1$ **to** E_{global} **do**
    $W_{central} = 0$; $S_{raw} = []$
    Let $\Theta'_{edge} = []$ be a list to store edge model weights
    for $i = 1$ **to** N **do**
      Initialize edge model M_i with $\theta_i = \theta_c$
      for $e = 1$ **to** E **do**
        Train M_i on D_i (batch B, learning rate lr)
      end for

      Compute a_i (accuracy of M_i on D_{val})
      Compute ϵ_i (prediction entropy of M_i on D_{val})
      $inv\_\epsilon_i = 1/(\epsilon_i + \delta_c)$
      $inv\_\epsilon_{cap} = \min(inv\_\epsilon_i, C_{max})$
      $u_i = (inv\_\epsilon_{cap} - inv\_\epsilon_{min})/(C_{max} - inv\_\epsilon_{min} + \epsilon_s)$
      $u_i = \max(0, \min(1, u_i))$ {Clamp to $[0, 1]$}
      $s_i = \lambda \cdot a_i + (1 - \lambda) \cdot u_i$
      Append s_i to S_{raw}
      Obtain updated weights θ'_i from M_i; Append θ'_i to Θ'_{edge}

    end for

    $Sum\_S_{raw} = \sum_{j=1}^{N} S_{raw}[j]$
    Let $S_{norm} = []$ be a list for normalized scaling factors
    for $i = 1$ **to** N **do**
      $S_{norm}[i] = S_{raw}[i]/(Sum\_S_{raw} + \epsilon_s)$
    end for

    for $i = 1$ **to** N **do**
      $\theta_{scaled_i} = S_{norm}[i] \cdot \Theta'_{edge}[i]$
      $W_{central} = W_{central} + \theta_{scaled_i}$
    end for

    $\theta_c = W_{central}$ {Update central model weights}
    Update M_c with θ_c
  end for

  return central model M_c
end function

Chapter 5

Result Analysis

5.1 Centralized Learning Results

5.1.1 Image-Only Classification on HAM10000

*All best-performing results for each model and configuration are highlighted in **bold** within table 5.1 to put emphasize on the top-performing setups across architectures.

Model	Reference	Accuracy	AUC-ROC	MCC	F1 Macro
VGG16	[5]	0.802	0.892	0.518	0.508
VGG19	[5]	0.764	0.879	0.474	0.483
ResNet18	[3]	0.764	0.911	0.497	0.523
ResNet50	[3]	0.766	0.916	0.552	0.549
InceptionV3	[1]	0.742	0.896	0.477	0.464
EfficientNet-B0	[16]	0.745	0.894	0.522	0.522
EfficientNet-B3	[16]	0.745	0.909	0.499	0.488
EfficientNet-B7	[16]	0.681	0.861	0.424	0.395
ViT-B/16	[17]	0.772	0.918	0.558	0.556
Swin Transformer-S	[19]	**0.781**	0.924	**0.580**	**0.571**
ConvNeXt-T	[24]	0.780	**0.928**	0.578	0.570

Table 5.1: Best performance comparison of CNN and Transformer-based architectures.

Classification Performance Analysis

VGG Architectures (VGG16 and VGG19)

The VGG-based models demonstrated moderate performance. VGG16 [5] achieved the highest performance within this group, with an accuracy of **0.802** and a weighted

F1-score of **0.743**. In comparison, VGG19 [5] achieved an accuracy of 0.764 and an AUC-ROC of 0.879. Overall, the simpler VGG16 architecture generalized better than VGG19 on the given dataset, indicating that increased depth did not provide a clear performance advantage.

ResNet Architectures (ResNet18 and ResNet50)

ResNet architectures consistently outperformed the VGG models. ResNet18 [3] achieved balanced performance with an accuracy of 0.764 and an AUC-ROC of 0.911.
ResNet50 [3] emerged as the strongest CNN-based model, achieving an accuracy of **0.766**, an AUC-ROC of **0.916**, and the highest MCC among CNNs (**0.552**). This highlights the effectiveness of residual connections in learning robust feature representations.

InceptionV3

InceptionV3 [1] achieved competitive performance with an AUC-ROC of 0.896; however, its macro F1-score remained relatively low at 0.464. This suggests that while InceptionV3 performed well in overall classification, it struggled with balanced class-wise predictions, particularly for minority classes.

EfficientNet Family (B0, B3, and B7)

The EfficientNet [16] models demonstrated consistent performance across variants. EfficientNet-B0 [20] and EfficientNet-B3 [20] both achieved an accuracy of 0.745, with EfficientNet-B0 attaining a strong MCC of **0.522**. EfficientNet-B3 recorded an AUC-ROC of 0.909, indicating reliable discriminative capability. EfficientNet-B7 [20] showed comparatively lower performance, with an accuracy of 0.681 and an AUC-ROC of 0.861, suggesting possible overfitting due to increased model complexity.

Transformer-Based Architectures

Transformer-based models generally outperformed CNN architectures. ViT-B/16 [17] achieved an accuracy of 0.772 and an AUC-ROC of 0.918.

Swin Transformer-S [19] achieved the best overall performance among all evaluated models, with an accuracy of **0.781**, an AUC-ROC of **0.924**, an MCC of **0.580**, and the highest macro and weighted F1-scores (**0.571** and **0.764**, respectively).

ConvNeXt-T [24] also performed competitively, achieving the highest AUC-ROC of **0.928**, while maintaining strong performance across other metrics.

Overall Comparison

Overall, Transformer-based architectures demonstrated superior performance in terms of classification robustness and class balance. While ResNet50 [3] remains the strongest CNN baseline, Swin Transformer-S [19] emerged as the most effective model, offering the best trade-off between accuracy and balanced classification performance. This trend is visually summarized in Figure ??, where Swin Transformer-S achieves the highest Macro-F1 score, and is further corroborated by the corresponding confusion matrix shown in Figure 5.2, which illustrates improved class-wise prediction consistency.

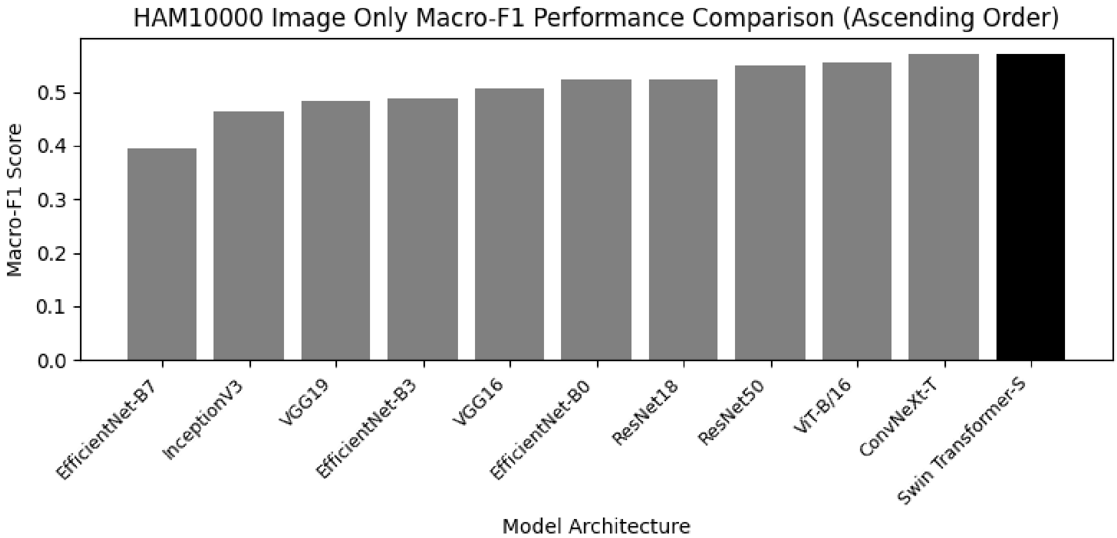


Figure 5.1: Performance Comparison of Macro-F1 on the HAM10000 Dataset

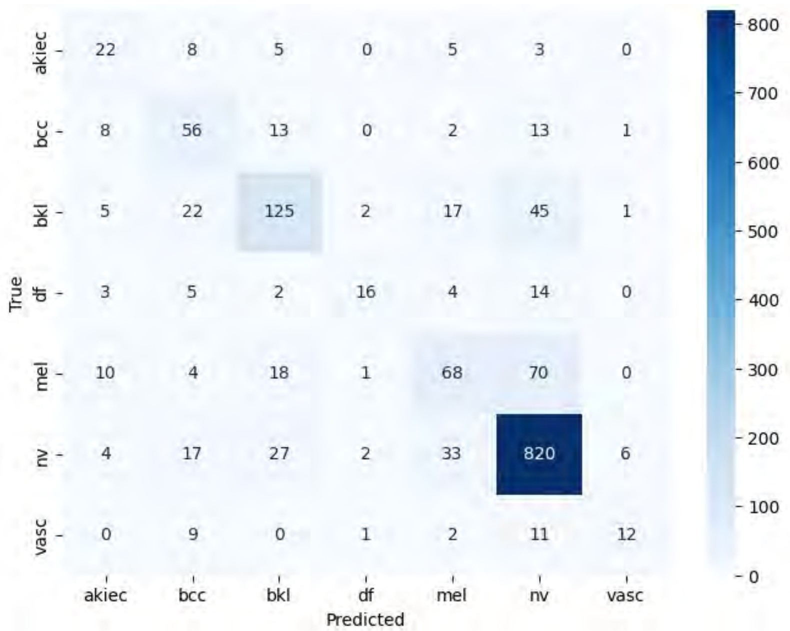


Figure 5.2: Confusion Matrix of Swin Transformer-S

5.1.2 Image-Only Classification on MILK10k

This experiment evaluates the generalization ability of image-only classification models across clinical and dermoscopic modalities using the MILK10k [48] dataset.

All models are trained under identical settings to ensure fair comparison.

Clinical-only Swin

As shown in Table 5.2, Swin-Small [19] achieves the highest overall performance among evaluated architectures on clinical images. It outperforms convolutional baselines in terms of Macro F1-score, AUC-ROC, and MCC, indicating superior robustness under class imbalance.

Model	**Reference**	**Accuracy**	**Macro-F1**	**AUC (Macro)**	**MCC**
ResNet50	[3]	0.8589	0.2856	0.8114	0.4648
ResNetV2-50	[8]	0.8350	0.2623	0.8059	0.4315
ResNetV2-101	[8]	0.8190	0.2735	0.8217	0.4066
InceptionV3	[1]	0.8371	0.3060	0.8419	0.4208
Inception-ResNetV2	[1]	0.8663	0.3069	0.8333	0.4902
MobileNetV2	[10]	0.8685	0.3071	0.8236	0.4836
DenseNet121	[11]	0.8840	0.3209	0.8330	0.5126
DenseNet169	[11]	0.8898	0.3444	0.8440	0.5397
DenseNet201	[11]	0.8966	0.3434	0.8293	0.5600
EfficientNet-B0	[16]	0.8462	0.2851	0.7987	0.4582
EfficientNet-B1	[16]	0.8691	0.3111	0.8208	0.4898
EfficientNet-B2	[16]	0.9060	0.3518	0.8022	0.5437
EfficientNet-B3	[16]	0.8838	0.3302	0.8171	0.5232
EfficientNet-B4	[16]	0.8886	0.3453	0.8105	0.5014
EfficientNetV2-B0	[20]	0.8873	0.3243	0.8107	0.5223
EfficientNetV2-B1	[20]	0.8916	0.3309	0.8106	0.5361
EfficientNetV2-B2	[20]	0.8835	0.3302	0.8330	0.5106
EfficientNetV2-B3	[20]	0.8907	0.3455	0.8409	0.5070
EfficientNetV2-S	[20]	0.9062	0.3578	0.8442	0.5495
ConvNeXt-Small	[24]	0.8348	0.2622	0.7972	0.4336
ViT-B/16	[17]	0.7749	0.2463	0.7807	0.3640
ViT-B/32	[17]	0.8307	0.2895	0.8324	0.4227
ViT-S/16	[17]	0.8796	0.3055	0.8191	0.4994
Swin-Tiny	[19]	0.8570	0.2924	0.8371	0.4742
Swin-Small	[19]	**0.9150**	**0.3764**	**0.8656**	**0.5945**
Swin-Base	[19]	0.8779	0.3146	0.8062	0.5146

Table 5.2: Comprehensive clinical image-only classification performance comparison on the MILK10k dataset.

This performance trend is further illustrated in Figure 5.3, where Swin-Small achieves the highest Macro-F1 score among all evaluated clinical-only models.

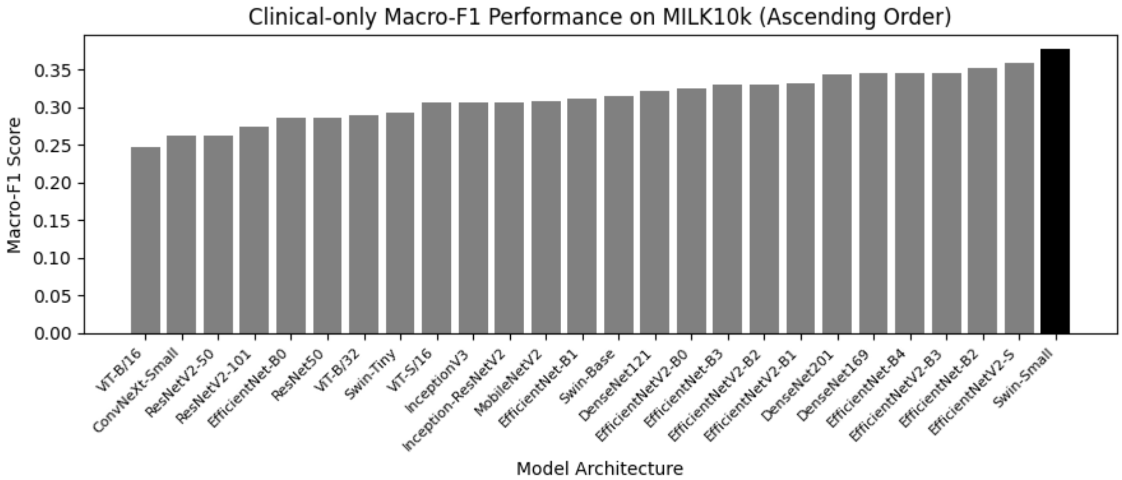


Figure 5.3: Macro-F1 score comparison of clinical image-only classification models on the MILK10k dataset

Dermoscopic-only Swin

Table 5.3 reports performance on dermoscopic images. Swin-Small [19] again achieves the best overall results, demonstrating consistent generalization across imaging modalities and validating its selection as the primary visual backbone for subsequent multimodal and federated experiments.

This performance trend is further illustrated in Figure 5.4, where Swin-Small achieves the highest Macro-F1 score among all dermoscopic-only models.

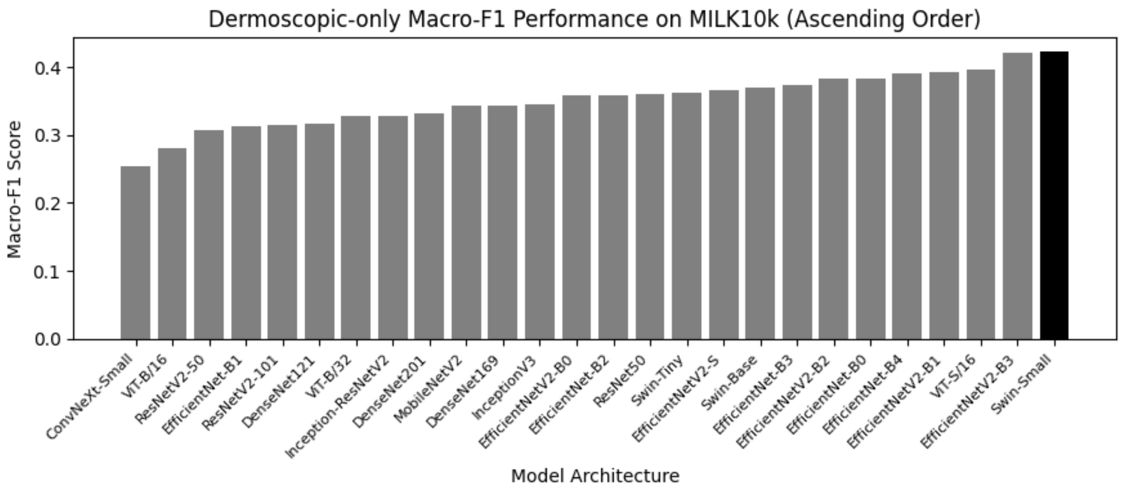
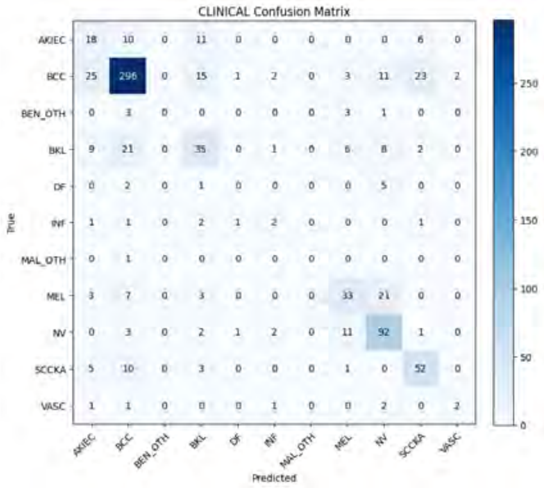


Figure 5.4: Macro-F1 score comparison of dermoscopic image-only classification models on the MILK10k dataset

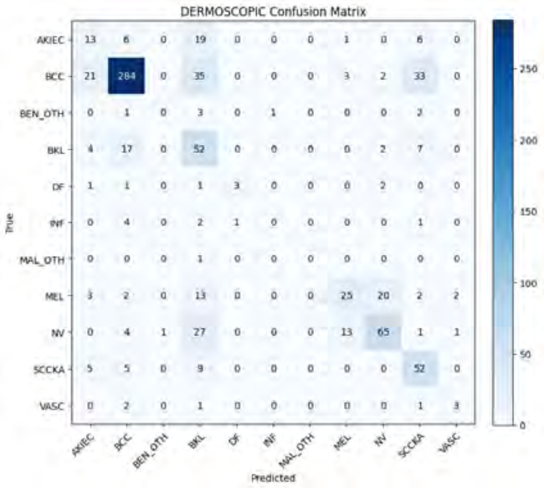
Model	Reference	Accuracy	Macro-F1	AUC (Macro)	MCC
ResNet50	[3]	0.8895	0.3598	0.8770	0.5507
ResNetV2-50	[8]	0.8492	0.3078	0.8224	0.4700
ResNetV2-101	[8]	0.8574	0.3150	0.8500	0.4845
InceptionV3	[1]	0.8812	0.3460	0.8477	0.5367
Inception-ResNetV2	[1]	0.8802	0.3290	0.8153	0.5105
MobileNetV2	[10]	0.8838	0.3435	0.8296	0.5310
DenseNet121	[11]	0.8602	0.3174	0.8275	0.4847
DenseNet169	[11]	0.8627	0.3437	0.8771	0.4987
DenseNet201	[11]	0.8763	0.3328	0.8536	0.5287
EfficientNet-B0	[16]	0.9028	0.3832	0.8522	0.5508
EfficientNet-B1	[16]	0.8661	0.3135	0.8394	0.5029
EfficientNet-B2	[16]	0.8820	0.3591	0.8654	0.5192
EfficientNet-B3	[16]	0.8941	0.3740	0.8773	0.5495
EfficientNet-B4	[16]	0.9204	0.3913	0.8389	0.6098
EfficientNetV2-B0	[20]	0.9043	0.3576	0.8486	0.5770
EfficientNetV2-B1	[20]	0.9214	0.3923	0.8712	0.6090
EfficientNetV2-B2	[20]	0.9159	0.3825	0.8449	0.5870
EfficientNetV2-B3	[20]	0.9217	0.4209	0.8914	0.6194
EfficientNetV2-S	[20]	0.8824	0.3657	0.8615	0.5354
ConvNeXt-Small	[24]	0.8005	0.2537	0.7492	0.4012
ViT-B/16	[17]	0.8434	0.2811	0.7878	0.4524
ViT-B/32	[17]	0.8573	0.3275	0.8100	0.4857
ViT-S/16	[17]	0.9003	0.3956	0.8841	0.5682
Swin-Tiny	[19]	0.8892	0.3631	0.8577	0.5506
Swin-Small	[19]	**0.9277**	**0.4221**	**0.8761**	**0.6558**
Swin-Base	[19]	0.8984	0.3695	0.8798	0.5828

Table 5.3: Comprehensive dermoscopic image-only classification performance on the MILK10k dataset.

Figure 5.5 presents the confusion matrices for clinical and dermoscopic image classification.



(a) Clinical image confusion matrix.



(b) Dermoscopic image confusion matrix.

Figure 5.5: Confusion matrices illustrating class-wise prediction behavior across clinical and dermoscopic imaging modalities.

5.1.3 HAM10000 Segmentation

The **U-Net** [4] model demonstrated strong performance in outlining lesion boundaries, and it has a effective balance of precision and recall. Its skip-connection architecture preserved fine spatial details which helped predict accurate segmentation of lesion regions even in low-contrast cases.

The quantitative results are shown below:

Metric	Value
IoU	0.8054
Dice	0.8753
Accuracy	0.9366
Precision	0.8934
Recall	0.8730

The U-Net achieved an IoU of 0.8054 and a Dice coefficient of 0.8753 which indicates a high degree of overlap between predicted and ground-truth lesion regions. Figure 5.6 shows the segmentation results using the U-Net model.

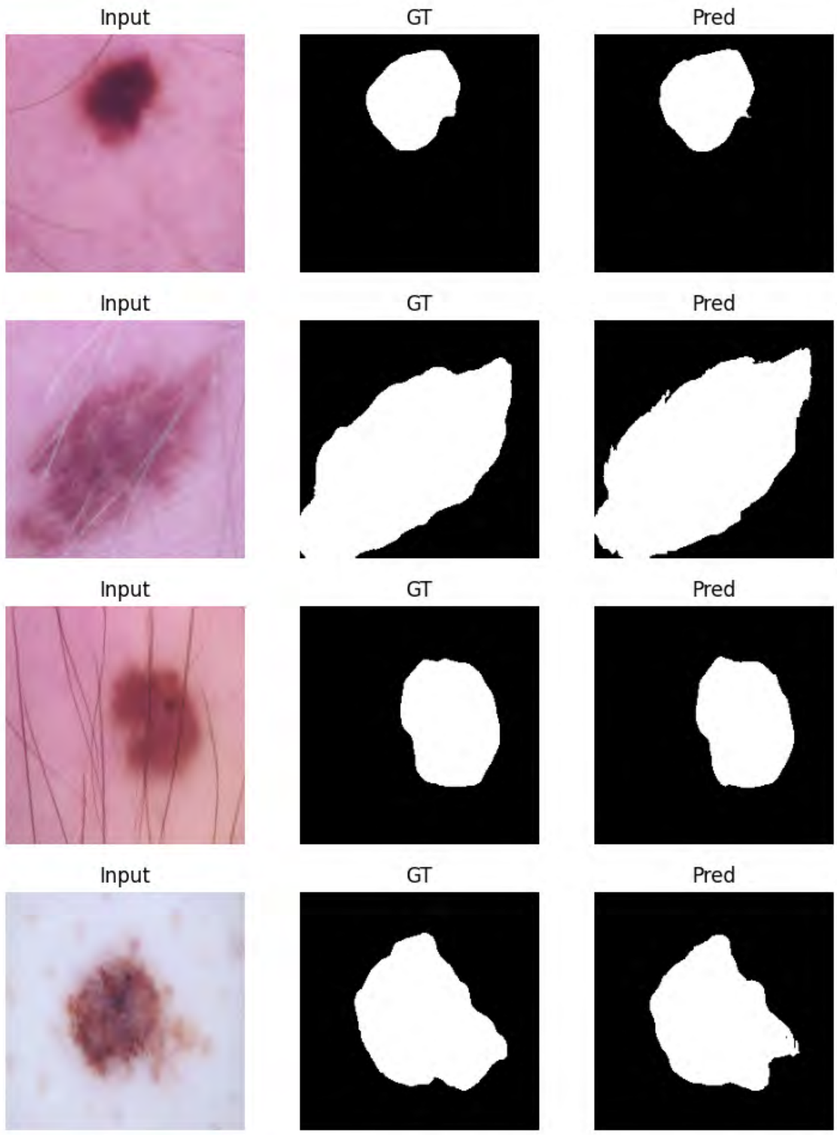


Figure 5.6: Qualitative segmentation results using the U-Net model.

The **Attention U-Net** [12] model outperformed the baseline U-Net by incorporating attention gates that selectively emphasized relevant lesion areas while suppressing background noise. This allowed for better localization of lesion boundaries, especially in complex cases with irregular shapes or illumination variations.
The model's performance metrics are as follows:

Metric	**Value**
IoU	0.8793
Dice	0.9295
Accuracy	0.9625
Precision	0.9425
Recall	0.9189

Compared to the standard U-Net [4], the Attention U-Net [12] achieved significant improvements across all key metrics, particularly in **IoU (+0.074)** and **Dice (+0.054)**. So it demonstrates a superior segmentation quality and boundary precision. Figure 5.7 shows the segmentation results using the Attention U-Net model. [12]

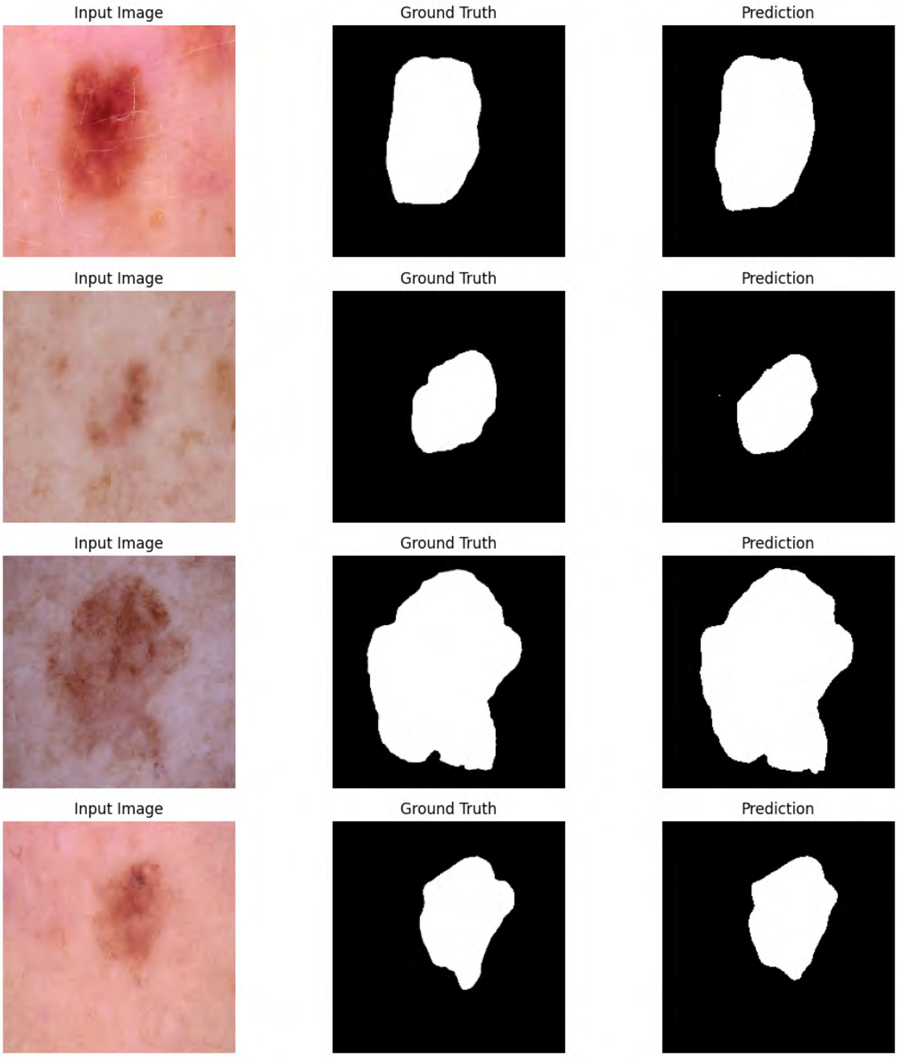


Figure 5.7: Qualitative segmentation results using the Attention U-Net model.

The **DeepLabV3** [9] model achieves the highest overall performance. It both U-Net [4] and Attention U-Net [12]. Its advantage primarily comes from the Atrous Spatial Pyramid Pooling (ASPP) module and deep ResNet backbone, which capture multiscale contextual information critical for precise lesion segmentation.

Metric	Value
IoU	0.9016
Dice	0.9444
Accuracy	0.9719
Precision	0.9651
Recall	0.9597

When compared to the baseline U-Net [4], the Attention U-Net [12], it shows consistent gains across all evaluation metrics, with approximately 9% higher IoU, 6% higher Dice, and moderate increases in accuracy (+2.8%), precision (+5.5%), and recall (+5.3%). Fig 5.8 shows segmentation results using the DeeplabV3 model.

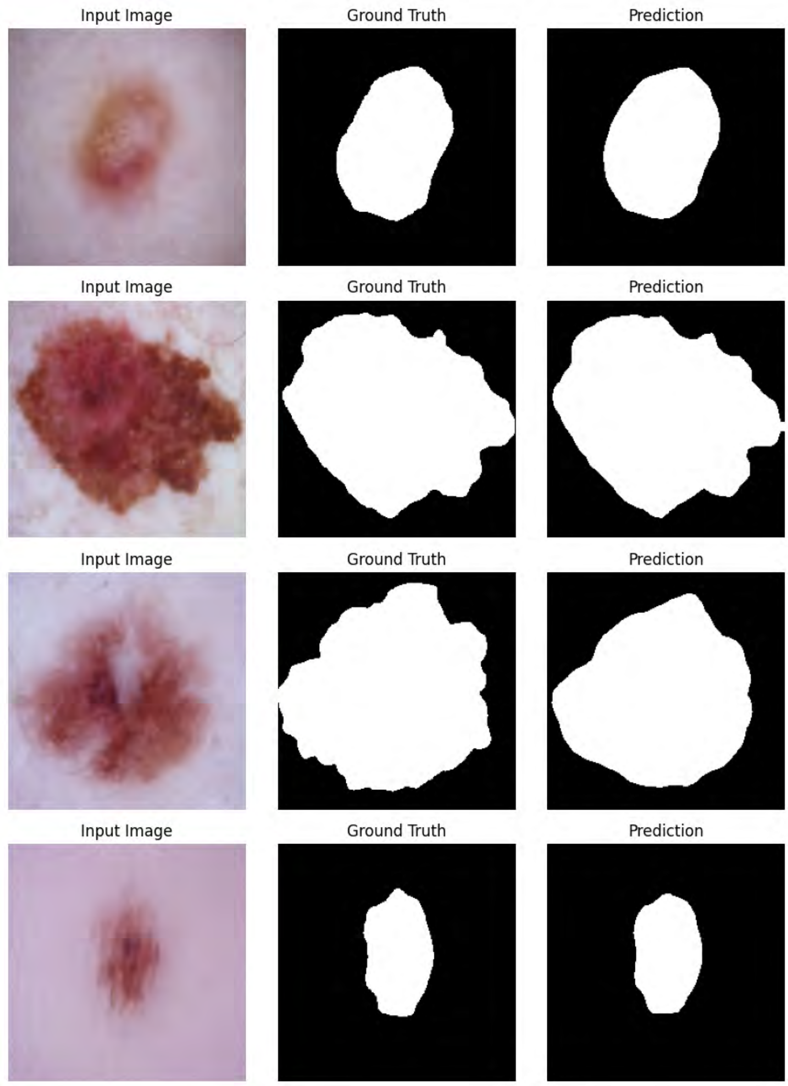


Figure 5.8: Qualitative segmentation results using the DeeplabV3 model.

Model	IoU	Dice	Pixel Accuracy	Precision	Recall
U-Net	0.8054	0.8753	0.9366	0.8934	0.8730
Attention U-Net	0.8793	0.9295	0.9625	0.9425	0.9189
DeepLabV3	**0.9016**	**0.9444**	**0.9719**	**0.9651**	**0.9597**

Table 5.4: Comparison of segmentation model performance metrics.

Table 5.4 presents a quantitative comparison of segmentation performance across U-Net, Attention U-Net, and DeepLabV3 models

5.1.4 Dual-Task Learning: Swin + Attention U-Net (HAM10000)

This experiment evaluates the impact of joint segmentation and classification learning using the proposed Swin [19] + Attention U-Net [12] architecture on the HAM10000 [14] dataset. In contrast to the image-only baseline presented in Section 5.1.1, this configuration introduces pixel-level lesion supervision as an auxiliary task, while maintaining the same Swin-based visual backbone.

The segmentation branch provides spatial guidance to the shared encoder, encouraging attention toward diagnostically relevant lesion regions. Performance is evaluated using standard segmentation metrics alongside classification metrics to assess the effect of multi-task learning on both tasks. The quantitative results of the joint segmentation and classification evaluation are summarized in Table 5.5.

Segmentation Metric	Score
Intersection over Union (IoU)	0.793
Dice Coefficient	0.882
Pixel Accuracy	0.939
Precision	0.913
Recall	0.857
Classification Metric	**Score**
Accuracy	0.738
AUC–ROC	0.917
Matthews Correlation Coefficient (MCC)	0.517
Macro F1-score	0.509
Weighted F1-score	0.742

Table 5.5: Joint segmentation and classification performance of the proposed Swin [19] + Attention U-Net [12] model on the HAM10000 [14] dataset.

Compared to the image-only Swin baseline, the dual-task configuration demonstrates improved lesion localization while maintaining competitive classification performance, highlighting the regularizing effect of auxiliary segmentation supervision.

The class-wise prediction behavior of the classification branch is further illustrated by the confusion matrix shown in Figure 5.9.

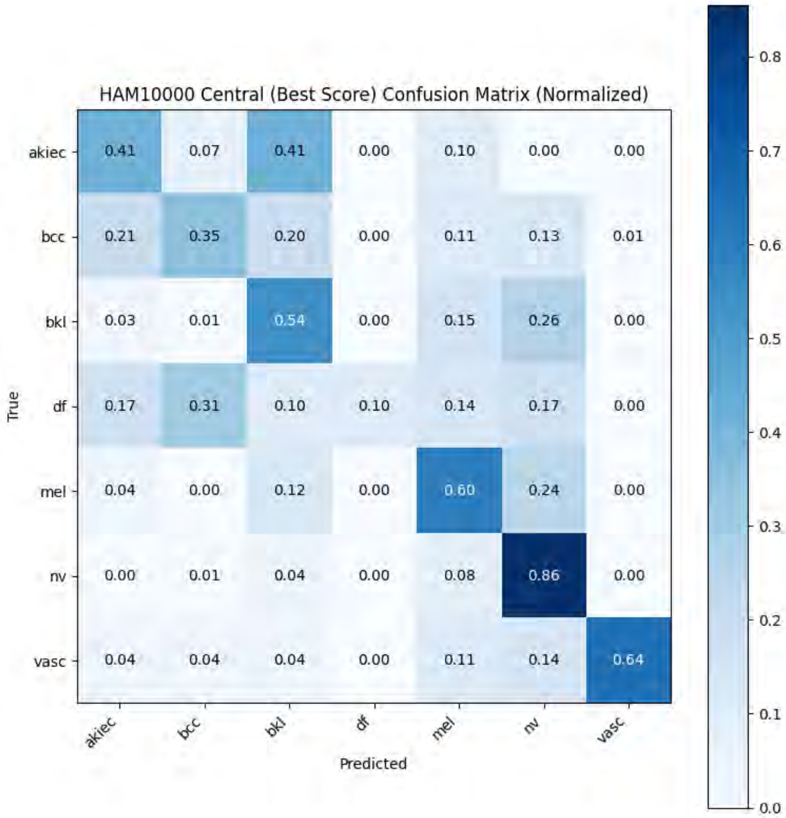
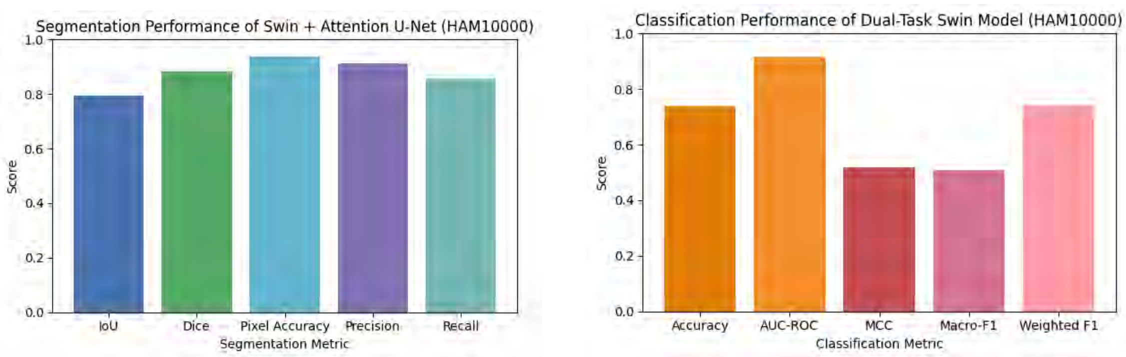


Figure 5.9: Confusion matrix of the classification branch of the dual-task Swin + Attention U-Net model on the HAM10000 dataset.

Figures 5.10 summarizes the segmentation and classification performance of the dual-task Swin + Attention U-Net model, respectively.



(a) Segmentation performance metrics. (b) Classification performance metrics.

Figure 5.10: Performance of the dual-task Swin + Attention U-Net model on the HAM10000 dataset.

To further analyze the training behavior of the proposed dual-task model, Fig. 5.11

illustrates the evolution of training and validation losses, while Fig. 5.12 reports validation segmentation Dice score and classification Macro-F1 across epochs.

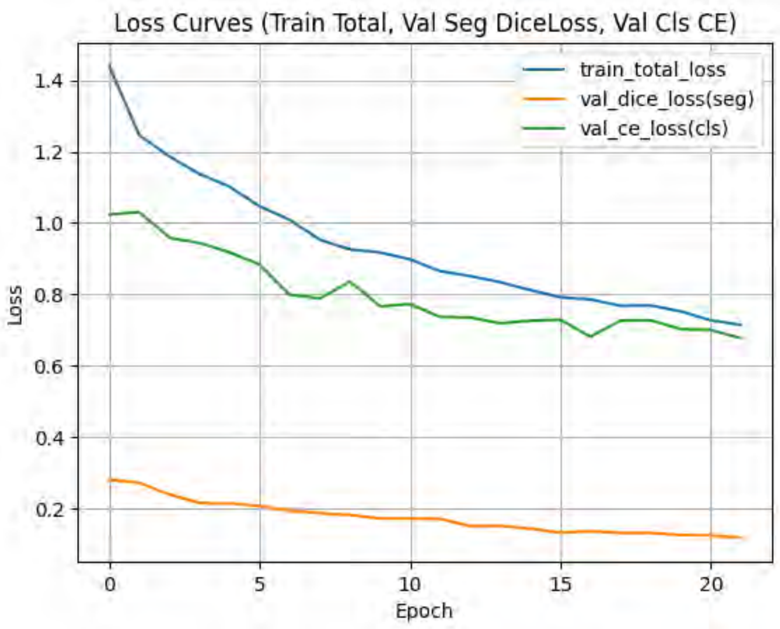


Figure 5.11: Training and validation loss curves of the dual-task Swin + Attention U-Net model on HAM10000.

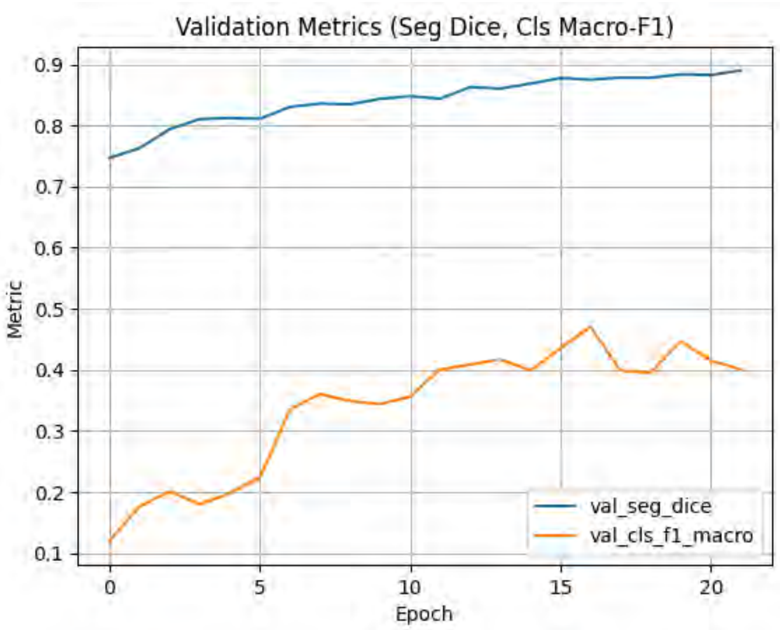


Figure 5.12: Validation segmentation Dice score and classification Macro-F1 score across training epochs on HAM10000.

5.1.5 Multimodal Fusion Results

HAM10000 + Metadata Fusion

This experiment evaluates the impact of incorporating patient metadata with dermoscopic image features on the HAM10000 [14] dataset. Three fusion strategies are considered: feature concatenation, gated fusion, and feature-wise linear modulation (FiLM).

As shown in Table 5.6, feature concatenation achieves the highest Macro F1-score and AUC-ROC, indicating strong overall classification performance. FiLM yields the highest balanced accuracy, suggesting improved class balance handling, while gated fusion provides competitive performance across all metrics.

The training and validation loss curves in Figure 5.13 further illustrate the convergence behavior of the different fusion strategies. While all approaches exhibit stable training dynamics, FiLM shows stronger late-stage overfitting, whereas concatenation and gated fusion demonstrate more consistent generalization across epochs.

These results demonstrate that metadata-aware fusion consistently improves classification performance over image-only baselines and motivates the integration of multimodal learning within the proposed federated framework.

Fusion Strategy	Accuracy	Macro F1	Balanced Acc.	AUC (Macro)
Concatenation	0.811	**0.644**	0.600	**0.946**
Gated Fusion	0.799	0.619	0.589	0.943
FiLM	0.803	0.637	**0.629**	0.935

Table 5.6: Multimodal fusion performance on the HAM10000 dataset using different fusion strategies.

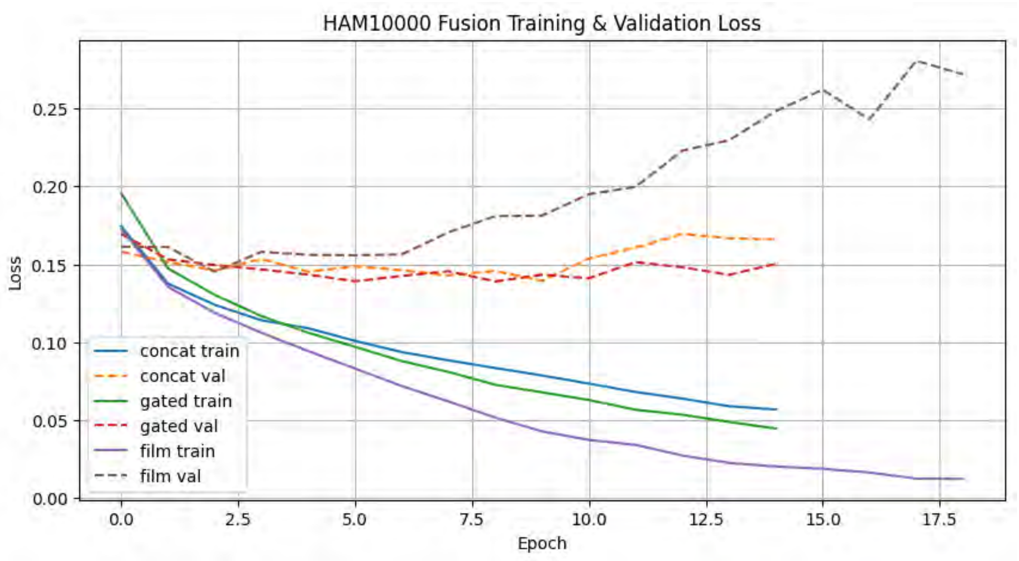


Figure 5.13: Training and Validation Loss Curve of HAM10000 fusion

MILK10k + Metadata Fusion

This experiment evaluates the effectiveness of different multimodal fusion strategies for integrating clinical and dermoscopic image features in a centralized setting. Four fusion mechanisms are considered: feature concatenation, gated fusion, and feature-wise linear modulation (FiLM).

As shown in Table 5.7, gated fusion achieves the highest Macro F1-score, indicating superior robustness under class imbalance. By explicitly learning adaptive weights to regulate the contribution of each modality, gated fusion enables more selective information integration compared to static fusion approaches.

The training and validation loss curves presented in Figure 5.14 illustrate the optimization behavior of the evaluated fusion strategies. Gated fusion exhibits more stable convergence and reduced divergence between training and validation loss, whereas FiLM demonstrates stronger overfitting tendencies in later epochs. Concatenation shows moderate generalization behavior but lacks the adaptive control observed in gated fusion.

Based on these results, gated fusion is selected as the preferred multimodal integration strategy for subsequent federated learning experiments.

Fusion Strategy	Accuracy	Macro F1	Balanced Acc.	AUC (Macro)
Concatenation	0.66	0.503	0.48	0.88
Gated Fusion	**0.68**	**0.519**	**0.51**	**0.90**
FiLM	0.64	0.479	0.46	0.87

Table 5.7: Multimodal fusion performance on the MILK10K dataset using different fusion strategies.

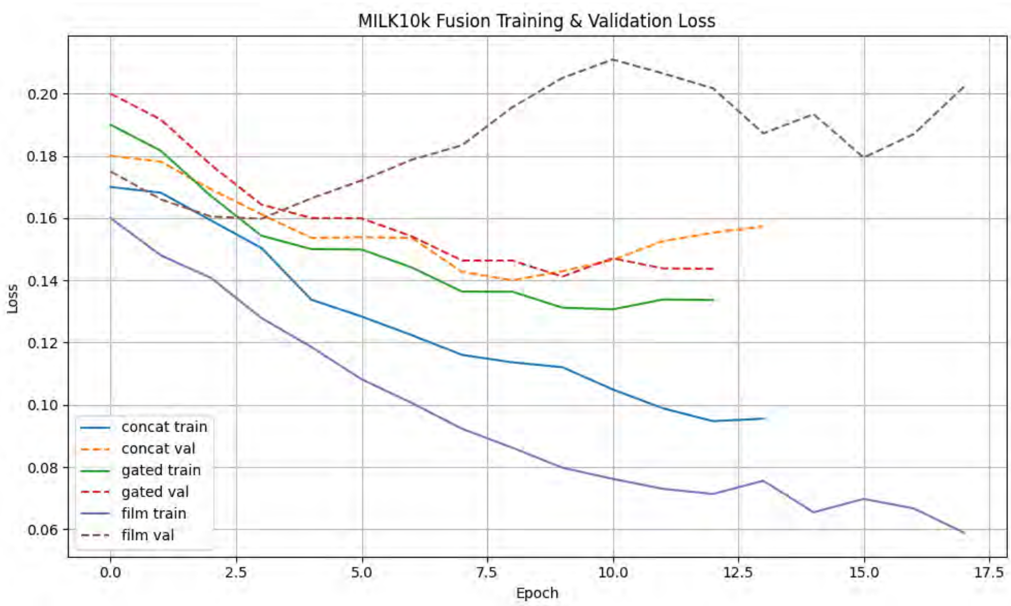


Figure 5.14: Training and Validation Loss Curve of MILK10K fusion

5.2 Federated Learning Results

5.2.1 Centralized vs Federated Performance (HAM10000)

On the HAM10000 [14] dataset, DUA-D2C [49] consistently preserves and in some cases improves classification performance compared to centralized training. As shown in Table 5.8, all fusion strategies exhibit positive performance gains under federated learning, with FiLM showing the largest improvement. This suggests that uncertainty-aware aggregation effectively mitigates client heterogeneity and suppresses unreliable updates, enabling robust global optimization.

Fusion Strategy	Centralized	Federated	Δ
Concatenation	0.676	0.684	+0.008
Gated Fusion	0.689	0.703	+0.014
FiLM	0.674	0.701	+0.027

Table 5.8: Centralized vs federated performance on HAM10000 [14] using DUA-D2C [49] aggregation (Macro F1-score).

5.2.2 Centralized vs Federated Performance (MILK10k)

On the MILK10k [48] dataset, federated learning introduces modest performance degradation due to increased data heterogeneity across clients. However, DUA-D2C [49] effectively constrains this degradation and enables selective performance recovery, particularly for FiLM-based fusion. These results indicate that uncertainty-aware aggregation improves robustness in complex multimodal federated environments.

Fusion Strategy	Centralized	Federated	Δ
Gated Fusion	0.494	0.471	-0.023
FiLM	0.443	0.468	+0.025
Concatenation	0.470	0.463	-0.007

Table 5.9: Centralized vs federated performance on MILK10k [48] using DUA-D2C [49] aggregation (Macro F1-score).

5.2.3 Fusion Strategy Behavior under Federated Learning

Across both HAM10000 [14] and MILK10k [48] datasets, distinct performance trends are observed among fusion strategies when transitioning from centralized to federated learning.

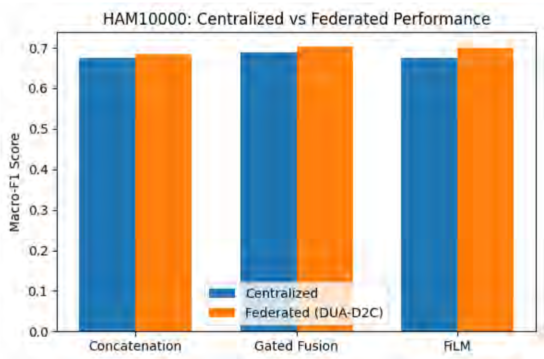
On the HAM10000 [14] dataset, all fusion strategies exhibit stable or improved performance under federated DUA-D2C [49] aggregation. In particular, gated fusion and FiLM demonstrate notable performance gains compared to centralized training. This behavior indicates that uncertainty-aware weighting effectively suppresses noisy client updates and enhances the contribution of reliable multimodal representations.

The relatively homogeneous nature of HAM10000 [14] further allows DUA-D2C [49] to exploit cross-client consensus, leading to improved global generalization.

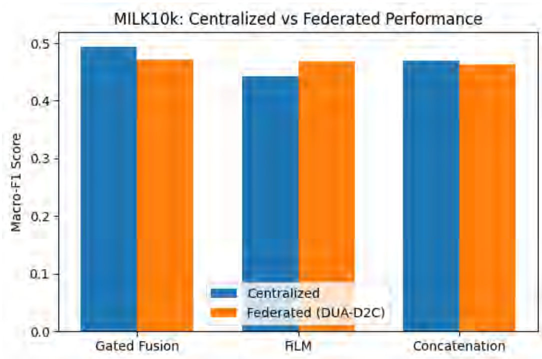
In contrast, MILK10k [48] presents a more challenging federated environment due to greater heterogeneity across clinical and dermoscopic modalities and higher inter-client variability. Under this setting, modest performance degradation is observed for some fusion strategies when moving to federated training. Nevertheless, DUA-D2C [49] consistently limits this degradation and, in the case of FiLM-based fusion, enables performance recovery beyond centralized baselines.

Gated fusion remains the most robust strategy across both centralized and federated regimes, owing to its ability to dynamically regulate the relative importance of image and metadata features. By learning modality-specific gates, the model effectively mitigates the impact of noisy or weak modalities at the client level. This property aligns well with the uncertainty-driven weighting mechanism of DUA-D2C [49] , resulting in stable federated performance.

Simpler fusion strategies such as feature concatenation consistently underperform adaptive mechanisms in federated settings, as they treat all modalities equally and lack explicit suppression of unreliable features under non-IID data distributions. On HAM10000 [14] dataset (Figure 5.15a), all fusion strategies exhibit stable or improved Macro-F1 performance under federated DUA-D2C [49] aggregation, reflecting effective mitigation of unreliable client updates in a relatively homogeneous environment. In contrast, MILK10k [48] (Figure 5.15b) demonstrates increased sensitivity to inter-client and cross-modal heterogeneity, where FiLM-based fusion consistently limits performance degradation and enables selective recovery beyond centralized baselines, highlighting the robustness of uncertainty-aware aggregation in complex federated scenarios.



(a) HAM10000: Centralized vs federated performance across fusion strategies.



(b) MILK10k: Centralized vs federated performance across fusion strategies.

5.2.4 Federated Performance Comparison: Central vs FedAvg vs DUA-D2C

Table 5.10 compares centralized training with federated learning using FedAvg [32] and the proposed DUA-D2C [49] strategy for the dual-task Swin [19] + Attention U-Net [12] model. While FedAvg already improves performance over centralized training, DUA-D2C consistently achieves the best Macro-F1, AUC–ROC, and MCC scores. This demonstrates that uncertainty-aware aggregation effectively reduces the influence of unreliable client updates and enhances global model quality under heterogeneous data distributions.

Training Paradigm	**Macro-F1**	**AUC–ROC**	**MCC**
Centralized Training	0.5092	0.9174	0.5167
FedAvg [32]	0.5339	0.9354	0.5417
DUA-D2C (Proposed) [49]	**0.5458**	**0.9361**	**0.5579**

Table 5.10: Federated classification performance comparison for the Swin + Attention U-Net model on HAM10000.

The comparative performance of centralized training, FedAvg and DUA-D2C is shown in Figure 5.16, where uncertainty-aware aggregation consistently yields superior Macro-F1 and MCC scores.

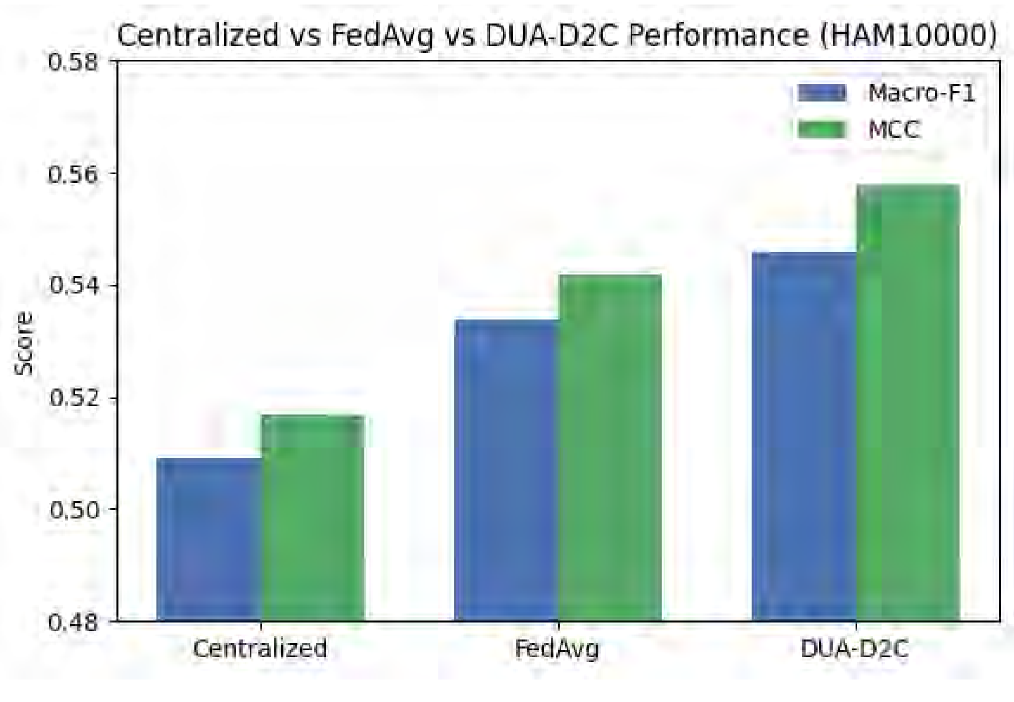


Figure 5.16: Comparison of centralized training, FedAvg, and the proposed DUA-D2C aggregation for the dual-task Swin + Attention U-Net model on HAM10000.

5.3 Discussion

5.3.1 Centralized Classification Performance

The centralized experiments first established strong baselines using image-only classification with Swin Transformer backbones. The results on HAM10000 [14] show that transformer-based architectures provide strong feature extraction capability for dermoscopic lesion images, capable of achieving higher macro-F1 scores compared to traditional CNN-based models. This indicates that Swin Transformer-S' [19] are effective at capturing both local lesion textures and global contextual patterns, which are important for distinguishing visually similar lesion categories.
However, performance degradation was observed when the same image-only models were evaluated on the more heterogeneous MILK10K [48] dataset. This indicates that classification robustness remains sensitive to variations in imaging modality, class imbalance and distribution shifts. These results suggest that while image-only learning can perform well in controlled settings, it may struggle to generalize across diverse clinical environments.

5.3.2 Impact of Dual-Branch Multi-Task Learning

The introduction of lesion segmentation as an auxiliary task significantly enhanced classification robustness by coupling Swin Transformer-S [19] feature extraction with an Attention U-Net [12] segmentation decoder. The results demonstrate that incorporating segmentation as an auxiliary task improves the quality of learned representations.
By explicitly modeling lesion boundaries, the segmentation branch guides the shared encoder to focus on diagnostically relevant regions rather than background artifacts. This auxiliary supervision enhances classification performance and improves feature robustness. Hence, the dual-task formulation provides a more clinically meaningful learning process, where lesion localization and diagnosis reinforce each other.

5.3.3 Multimodal Fusion Benefits

The multimodal fusion experiments further show that integrating patient metadata alongside image features consistently improves diagnostic performance compared to image-only baselines. Clinical attributes such as age, lesion site and skin tone provide complementary information that cannot always be inferred from visual appearance alone.
The results on both HAM10000 [14] and MILK10K [48] confirm that multimodal learning improves generalization and reduces misclassification in challenging lesion categories. This aligns with real-world dermatological practice, where physicians rely not only on images but also on contextual patient information for accurate diagnosis.

5.3.4 Federated Learning Performance and Challenges

In the federated learning setting, the baseline FedAvg [32] aggregation method enabled privacy-preserving collaboration across distributed clients without sharing raw medical data. However, the results indicate that FedAvg [32] is sensitive to non-IID client distributions, which are common in multi-institutional medical datasets. Client heterogeneity, varying data quality, and imbalance across lesion categories can lead to unstable convergence and suboptimal global models. These limitations motivate the need for adaptive aggregation strategies that account for client reliability.

5.3.5 Benefits of the Proposed DUA-D2C Aggregation

The proposed Dynamic Uncertainty-Aware Divide2Conquer (DUA-D2C) [49] aggregation strategy addresses these challenges by weighting client updates based on both predictive performance and uncertainty estimation. Clients producing unreliable or highly uncertain predictions or lower macro-F1 contribution are automatically down-weighted during global aggregation. This reduces the influence of unreliable or noisy client updates and improves the stability of federated optimization. The experimental comparisons show that DUA-D2C [49] preserves or improves performance on HAM10000 [14] while constraining degradation on the more heterogeneous MILK10K [48] dataset. Hence, uncertainty-aware aggregation provides a more robust and reliability-driven federated learning mechanism, which is especially important for clinical deployment.

Chapter 6

Conclusion

6.1 Summary of Findings

This thesis investigated an uncertainty-aware federated dual-branch learning framework for privacy-preserving skin lesion diagnosis. The proposed approach was designed to overcome key limitations of centralized training, unimodal learning, and conventional federated aggregation when applied to heterogeneous medical imaging environments.

Comprehensive experiments were conducted on the HAM10000 [14] and MILK10k [48] datasets, covering image-only classification, dual-task learning with segmentation supervision, multimodal fusion with patient metadata, and federated learning under non-IID data distributions. Empirical results demonstrated that the Swin Transformer consistently outperformed convolutional and alternative transformer-based backbones across both dermoscopic and clinical modalities, validating its selection as the shared visual encoder.

For HAM10000 [14], joint optimization of classification and segmentation using a Swin + Attention U-Net architecture improved spatial localization and representation robustness while maintaining strong diagnostic performance. For MILK10k, multimodal fusion of image features with structured metadata further enhanced classification accuracy and macro-averaged F1 scores compared to unimodal baselines.

In federated settings, conventional FedAvg [32] aggregation was found to be sensitive to client heterogeneity and predictive uncertainty. The proposed Dynamic Uncertainty-Aware Divide2Conquer (DUA-D2C) [49] aggregation strategy addressed these limitations by integrating client performance and uncertainty into the aggregation process, resulting in more stable and reliable global models. Overall, the findings confirm the effectiveness of uncertainty-aware dual-branch multimodal learning for distributed medical image analysis.

6.2 Contributions to the Field

The main contributions of this thesis are summarized as follows:

- A unified dual-branch learning framework that integrates task-level supervision (classification and segmentation) and modality-level learning (image and metadata) within a single, modular architecture.
- A systematic evaluation and validation of the Swin Transformer [19] as a robust visual backbone for dermoscopic and clinical skin lesion analysis.
- The introduction of a Dynamic Uncertainty-Aware Divide2Conquer (DUA-D2C) [49] federated aggregation strategy that adaptively weights client updates based on predictive performance and uncertainty.
- An extensive experimental study covering centralized and federated learning scenarios on HAM10000 [14] and MILK10k [48] , including image-only, dual-task, and multimodal fusion settings.
- Practical insights into the behavior of multimodal fusion strategies under both centralized and federated learning paradigms.

These contributions collectively advance the robustness, interpretability, and privacy-preserving capabilities of federated deep learning systems for medical image analysis.

6.3 Recommendations for Future Work

While the proposed framework demonstrates strong performance across centralized and federated settings, several promising research directions remain open for future investigation.To begin with, the future work can consider the personalized federated learning strategies, which will enable. partial client-level model adaptation. This may facilitate such ways more. normalize extreme data heterogeneity between institutions, through global knowledge balancing. sharing with local expertise especially with clinically heterogenous populations. Second, it is possible to further develop uncertainty modeling by including Bayesian. deep learning methods, including Monte Carlo dropout [7], variational inference, or ensemble-based uncertainty estimation. Such approaches can be more credible. confidence estimates which are essential to high stakes clinical decision-making and can also encourage the suggested uncertainty-conscious aggregation approach. Third, the generalization of the framework to ongoing and continuous instances of learning. would allow the system to be altered to new lesion cate, changing distributions of data. gories, or modifications to acquisition procedures without the need to completely retrain. This capability is particularly applicable in actual medical deployments in the real world where data char is needed. The traits are not fixed. Also, further research may focus on larger-scale federated deployments in. Involving more clients, more various imaging tools, and cross-domain. datasets. These kinds of assessments would give more information on scalability, communi. cation efficiency and resiliency against realistic clinical network behavior.

Bibliography

[1] C. Szegedy et al., “Going deeper with convolutions,” *CoRR*, vol. abs/1409.4842, 2014. arXiv: 1409.4842. [Online]. Available: http://arxiv.org/abs/1409.4842

[2] E. Wilmer, C. Gustafson, C. Ahn, S. Davis, S. Feldman, and W. Huang, “Most common dermatologic conditions encountered by dermatologists and nondermatologists,” *Cutis*, vol. 94, pp. 285–92, Dec. 2014.

[3] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2015. arXiv: 1512.03385 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/1512.03385

[4] O. Ronneberger, P. Fischer, and T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, 2015. arXiv: 1505.04597 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/1505.04597

[5] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 2015. arXiv: 1409.1556 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/1409.1556

[6] R. Sumithra, M. Suhil, and D. Guru, “Segmentation and classification of skin lesions for disease diagnosis,” *Procedia Computer Science*, vol. 45, pp. 76–85, 2015, International Conference on Advanced Computing Technologies and Applications (ICACTA), ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2015.03.090 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050915003269

[7] Y. Gal and Z. Ghahramani, *Dropout as a bayesian approximation: Representing model uncertainty in deep learning*, 2016. arXiv: 1506.02142 `[stat.ML]`. [Online]. Available: https://arxiv.org/abs/1506.02142

[8] K. He, X. Zhang, S. Ren, and J. Sun, *Identity mappings in deep residual networks*, 2016. arXiv: 1603.05027 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/1603.05027

[9] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, *Rethinking atrous convolution for semantic image segmentation*, 2017. arXiv: 1706.05587 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/1706.05587

[10] A. G. Howard et al., *Mobilenets: Efficient convolutional neural networks for mobile vision applications*, 2017. arXiv: 1704.04861 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/1704.04861

[11] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, *Densely connected convolutional networks*, 2018. arXiv: 1608.06993 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/1608.06993

[12] O. Oktay et al., *Attention u-net: Learning where to look for the pancreas*, 2018. arXiv: 1804.03999 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/1804.03999

[13] A. Rezvantalab, H. Safigholi, and S. Karimijeshni, *Dermatologist level dermoscopy skin cancer classification using different deep learning convolutional neural networks algorithms*, 2018. arXiv: 1810.10348 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/1810.10348

[14] P. Tschandl, "The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Harvard Dataverse*, vol. V4, 2018, Dataset. DOI: 10.7910/DVN/DBW86T [Online]. Available: https://doi.org/10.7910/DVN/DBW86T

[15] J. R. Hagerty et al., "Deep learning and handcrafted method fusion: Higher diagnostic accuracy for melanoma dermoscopy images," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1385–1391, 2019. DOI: 10.1109/JBHI.2019.2891049

[16] M. Tan and Q. V. Le, *Efficientnet: Rethinking model scaling for convolutional neural networks*, 2020. arXiv: 1905.11946 `[cs.LG]`. [Online]. Available: https://arxiv.org/abs/1905.11946

[17] A. Dosovitskiy et al., *An image is worth 16x16 words: Transformers for image recognition at scale*, 2021. arXiv: 2010.11929 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/2010.11929

[18] S. Jain, U. Singhania, B. Tripathy, E. A. Nasr, M. K. Aboudaif, and A. K. Kamrani, "Deep learning-based transfer learning for classification of skin cancer," *Sensors*, vol. 21, no. 23, 2021, ISSN: 1424-8220. DOI: 10.3390/s21238142 [Online]. Available: https://www.mdpi.com/1424-8220/21/23/8142

[19] Z. Liu et al., *Swin transformer: Hierarchical vision transformer using shifted windows*, 2021. arXiv: 2103.14030 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/2103.14030

[20] M. Tan and Q. V. Le, *Efficientnetv2: Smaller models and faster training*, 2021. arXiv: 2104.00298 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/2104.00298

[21] H. Q. Yu and S. Reiff-Marganiec, "Targeted ensemble machine classification approach for supporting iot enabled skin disease detection," *IEEE Access*, vol. 9, pp. 50 244–50 252, 2021. DOI: 10.1109/ACCESS.2021.3069024

[22] E. K. Aghdam, R. Azad, M. Zarvani, and D. Merhof, "Attention swin u-net: Cross-contextual attention mechanism for skin lesion segmentation," *arXiv preprint arXiv:2210.16898*, 2022. [Online]. Available: https://arxiv.org/abs/2210.16898

[23] W. Gouda, N. Sama, G. Al-Waakid, M. Humayun, and N. Jhanjhi, "Detection of skin cancer based on skin lesion images using deep learning," *Healthcare*, vol. 10, p. 1183, Jun. 2022. DOI: 10.3390/healthcare10071183

[24] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, *A convnet for the 2020s*, 2022. arXiv: 2201.03545 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/2201.03545

[25] R. A. Mehr and A. Ameri, "Skin cancer detection based on deep learning," *Journal of Biomedical Physics & Engineering*, vol. 12, no. 6, pp. 559–568, Dec. 2022, ISSN: 2251-7200. DOI: 10.31661/jbpe.v0i0.2207-1517 [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9759648/

[26] C. Ou et al., "A deep learning based multimodal fusion model for skin lesion diagnosis using smartphone collected clinical images and metadata," *Frontiers in Surgery*, vol. Volume 9 - 2022, 2022, ISSN: 2296-875X. DOI: 10.3389/fsurg.2022.1029991 [Online]. Available: https://www.frontiersin.org/journals/surgery/articles/10.3389/fsurg.2022.1029991

[27] S. Wang, Y. Yin, D. Wang, Y. Wang, and Y. Jin, "Interpretability-based multimodal convolutional neural networks for skin lesion diagnosis," *IEEE Transactions on Cybernetics*, vol. 52, no. 12, pp. 12 623–12 637, 2022. DOI: 10.1109/TCYB.2021.3069920

[28] H. L. Gururaj, N. Manju, A. Nagarjun, V. N. M. Aradhya, and F. Flammini, "Deepskin: A deep learning approach for skin cancer classification," *IEEE Access*, vol. 11, pp. 50 205–50 214, 2023. DOI: 10.1109/ACCESS.2023.3274848

[29] M. H. Luu et al., "Quantification of liver-lung shunt fraction on 3d spect/ct images for selective internal radiation therapy of liver cancer using cnn-based segmentations and non-rigid registration," *Computer Methods and Programs in Biomedicine*, vol. 233, p. 107 453, 2023, ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2023.107453 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260723001190

[30] P. A. Lyakhov, U. A. Lyakhova, and D. I. Kalita, "Multimodal analysis of unbalanced dermatological data for skin cancer recognition," *IEEE Access*, vol. 11, pp. 131 487–131 507, 2023. DOI: 10.1109/ACCESS.2023.3336289

[31] P. A. Lyakhov, U. A. Lyakhova, and D. I. Kalita, "Multimodal analysis of unbalanced dermatological data for skin cancer recognition," *IEEE Access*, vol. 11, pp. 131 487–131 507, 2023. DOI: 10.1109/ACCESS.2023.3336289

[32] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, *Communication-efficient learning of deep networks from decentralized data*, 2023. arXiv: 1602.05629 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1602.05629

[33] B. Nam, "Skin disease classification using privacy-preserving federated learning," *International Journal of High School Research*, vol. 5, pp. 99–104, Feb. 2023. DOI: 10.36838/v5i1.19

[34] M. Tahir, A. Naeem, H. Malik, J. Tanveer, R. A. Naqvi, and S.-W. Lee, "DSCC_Net: Multi-Classification Deep Learning Models for Diagnosing of Skin Cancer Using Dermoscopic Images," *Cancers*, vol. 15, no. 7, 2023, ISSN: 2072-6694. DOI: 10.3390/cancers15072179 [Online]. Available: https://www.mdpi.com/2072-6694/15/7/2179

[35] A. Al Mahmud et al., "SkinNet-14: a deep learning framework for accurate skin cancer classification using low-resolution dermoscopy images with optimized training time," *Neural Computing and Applications*, vol. 36, no. 30, pp. 18 935–18 959, 2024, Published: 2024-10-01, ISSN: 1433-3058. DOI: 10.1007/s00521-024-10225-y [Online]. Available: https://doi.org/10.1007/s00521-024-10225-y

[36] C. Flosdorf, J. Engelker, I. Keller, and N. Mohr, *Skin cancer detection utilizing deep learning: Classification of skin lesion images using a vision transformer*, 2024. arXiv: 2407.18554 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2407.18554

[37] A. Gupta, M. M. Maurya, K. Dhere, and V. Chaurasiya, “Privacy-preserving hybrid federated learning framework for mental healthcare applications: Clustered and quantum approaches,” *IEEE Access*, vol. PP, pp. 1–1, Jan. 2024. DOI: 10.1109/ACCESS.2024.3464240

[38] E. H. Houssein, D. A. Abdelkareem, G. Hu, M. A. Hameed, I. A. Ibrahim, and M. Younan, “An effective multiclass skin cancer classification approach based on deep convolutional neural network,” *Cluster Computing*, vol. 27, no. 9, pp. 12 799–12 819, Dec. 2024, ISSN: 1573-7543. DOI: 10.1007/s10586-024-04540-1 [Online]. Available: https://doi.org/10.1007/s10586-024-04540-1

[39] M. Mateen, S. Hayat, F. Arshad, Y.-H. Gu, and M. A. Al-antari, “Hybrid deep learning framework for melanoma diagnosis using dermoscopic medical images,” *Diagnostics*, vol. 14, no. 19, 2024, ISSN: 2075-4418. DOI: 10.3390/diagnostics14192242 [Online]. Available: https://www.mdpi.com/2075-4418/14/19/2242

[40] A. Shah, S. Wahood, D. Guermazi, C. E. Brem, and E. Saliba, “Skin and syntax: Largedensenete models in dermatopathology,” *Dermatopathology*, vol. 11, no. 1, pp. 101–111, 2024, ISSN: 2296-3529. DOI: 10.3390/dermatopathology11010009 [Online]. Available: https://www.mdpi.com/2296-3529/11/1/9

[41] I. Wiest et al., “Privacy-preserving large language models for structured medical information retrieval,” *NPJ digital medicine*, vol. 7, p. 257, Sep. 2024. DOI: 10.1038/s41746-024-01233-2

[42] A. Yang and E. Yang, “A multimodal approach to the detection and classification of skin diseases,” *arXiv preprint arXiv:2411.13855*, 2024. [Online]. Available: https://arxiv.org/abs/2411.13855

[43] A. D. Khalaf, H. Hamdan, A. Abdul Halin, and N. Manshor, “Segmentation and classification of skin cancer diseases based on deep learning: Challenges and future directions,” *IEEE Access*, vol. 13, pp. 90 163–90 184, 2025. DOI: 10.1109/ACCESS.2025.3569170

[44] V. Khullar et al., “Minimal sourced and lightweight federated transfer learning models for skin cancer detection,” *Scientific Reports*, vol. 15, no. 1, p. 2605, Jan. 2025, ISSN: 2045-2322. DOI: 10.1038/s41598-024-82402-x [Online]. Available: https://doi.org/10.1038/s41598-024-82402-x

[45] J. Mohan, A. Sivasubramanian, S. V., and V. Ravi, “Enhancing skin disease classification leveraging transformer-based deep learning architectures and explainable ai,” *Computers in Biology and Medicine*, vol. 190, p. 110 007, May 2025, ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2025.110007 [Online]. Available: http://dx.doi.org/10.1016/j.compbiomed.2025.110007

[46] R. Panneerselvam, S. Balasubramaniam, V. Sureshkumar, V. Ravi, and S. Maidin, “Multimodal skin cancer prediction: Integrating dermoscopic images and clinical metadata with transfer learning,” *The Open Bioinformatics Journal*, vol. 18, Jan. 2025. DOI: 10.2174/0118750362358444250120070327

[47] N. Pham, P. Đat, L. Tan, and K. Huynh, "A multimodal deep ensemble framework for skin lesion classification," in Mar. 2025, pp. 100–111. DOI: 10.1007/978-981-96-4606-7_9

[48] T. Philipp et al., "Milk10k: A hierarchical multimodal imaging learning toolkit for diagnosing pigmented and non-pigmented skin cancer and its simulators," *Journal of Investigative Dermatology*, 2025. DOI: 10.1016/j.jid.2025.06.1594 [Online]. Available: https://doi.org/10.1016/j.jid.2025.06.1594

[49] M. S. B. Siddiqui, M. M. Islam, and M. G. R. Alam, *Dua-d2c: Dynamic uncertainty aware method for overfitting remediation in deep learning*, 2025. arXiv: 2411.15876 `[cs.LG]`. [Online]. Available: https://arxiv.org/abs/2411.15876

[50] S. Yan, X. Li, M. Hu, Y. Jiang, Z. Yu, and Z. Ge, "Make: Multi-aspect knowledge-enhanced vision-language pretraining for zero-shot dermatological assessment," *arXiv preprint arXiv:2505.09372*, 2025. [Online]. Available: https://arxiv.org/abs/2505.09372