SKIN LESION CLASSIFICATION USING DEEP LEARNING MODELS

PHASE II REPORT

Submitted by

ADITHAN K	**21011101009**
AGILAN MA	**21011101011**
C GANESH RAM	**21011101034**
DHEV MUGUNDDHAN A	**21011101040**

in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
ARTIFICIAL INTELLIGENCE
&
DATA SCIENCE

DEPARTMENT OF COMPUTER SCIENCE
AND ENGINEERING
SCHOOL OF ENGINEERING
SHIV NADAR UNIVERSITY CHENNAI

APRIL 2025

SHIV NADAR UNIVERSITY CHENNAI

BONAFIDE CERTIFICATE

Certified that this report titled "**SKIN LESION CLASSIFICATION USING DEEP LEARNING MODELS**" is the bonafide work of **ADITHAN K** (Reg. No: **21011101009**), **AGILAN MA** (Reg. No: **21011101011**), **C GANESH RAM** (Reg. No: **21011101034**), **DHEV MUGUNDDHAN A** (Reg. No: **21011101040**) who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr. T. Nagarajan

Professor & Head

Department of Computer Science & Engineering

School of Engineering

Shiv Nadar University Chennai,

Kalavakkam -603110.

SIGNATURE

Dr. Dhivya S

Assistant Professor

Department of Computer Science & Engineering

School of Engineering

Shiv Nadar University Chennai,

Kalavakkam -603110.

ABSTRACT

Skin cancer is one of the most prevalent and life-threatening diseases, requiring early and accurate diagnosis for effective treatment. This project aims to develop a robust, accurate, and scalable deep learning model utilizing the ISIC 2024 dataset to perform skin lesion classification. The dataset comprises dermoscopic images and patient metadata, presenting challenges such as class imbalance, variable image quality, multimodal input and missing data. To address these, advanced data processing techniques were used, such as imputation strategies, feature engineering, and class balancing methods.

State-of-the-art Deep Learning architectures were explored, including CNN-based models (ResNet, DenseNet), Vision Transformers (ViT, MedMamba, MobileViT), and multimodal architectures integrating metadata. Image segmentation was done using DeepLabv3+, enhancing lesion localization. A stacked ensemble model combining tree-based (XGBoost, LightGBM, CatBoost) and neural net classifiers was used to classify the images utilizing both Image and Metadata features.

The models were evaluated using the partial area under the ROC curve (pAUC) above 80% TPR, ensuring clinically relevant sensitivity. Experimental results indicate that multimodal architectures outperform single-modality models, and MedMamba, Vision Transformers, Densenet-169 models produced superior performance. The stacked ensemble approach further improved performance, reducing variance and increasing robustness against noisy data. This research contributes to the advancement of AI-assisted dermatology, improving the early detection and classification of skin cancer with enhanced accuracy and reliability.

Keywords - Skin lesion classification, ISIC 2024 challenge,MedMamba, ResNet, DenseNet, Vision Transformer, multimodal Deep Learning, ensemble learning.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to all those who have provided invaluable guidance, support, and encouragement throughout the course of this project.

First and foremost, We are deeply grateful to **Dr. T. Nagarajan**, Professor and Head, Department of Computer Science and Engineering, School of Engineering, Shiv Nadar University Chennai, for his steadfast guidance and unwavering encouragement. His insights into the field of artificial intelligence and cognitive science have been instrumental in shaping the direction of this research.

We would like to extend our heartfelt thanks to our supervisor, **Dr. Dhivya S**, Assistant Professor, Department of Computer Science and Engineering, School of Engineering, Shiv Nadar University Chennai, for her mentorship and commitment to the success of this project. Her expertise in the domain of Artificial Intelligence and Medical Imaging has been invaluable, providing us with the clarity and focus needed to tackle complex challenges and refine the methodology of this work.

We also acknowledge with gratitude the support of our family, whose understanding and encouragement have been crucial to the implementation of this project. Their unwavering belief in us has been a constant source of motivation and strength.

ADITHAN K
(21011101009)

AGILAN MA
(21011101011)

C GANESH RAM
(21011101034)

DHEV MUGUNDDHAN A
(21011101040)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS, ABBREVIATIONS

α_t	Class balancing factor for Focal Loss
γ	Focusing parameter for Focal Loss to handle class imbalance
$\hat{y}$	Predicted label (can be raw score in some cases)
p	Predicted probability of the positive class
p_t	Model's estimated probability for the true class
y	Ground truth label (0 or 1)
AI	Artificial Intelligence
CAD	Computer Aided Diagnostics
CATBOOST	Categorical Boosting
CNN	Convolutional Neural Networks
DeiT	Data-efficient image transformers
Ex-AI	Explainable AI
GAN	Generative Adversarial Networks
ISIC	International Skin Imaging Collaboration
MCMC	Markov Chain Monte Carlo
ResNet	Residual Networks
ROC	Receiver-Operating Characteristic
SMOTE	Synthetic minority oversampling technique

SMOTEENN	Synthetic minority oversampling technique and Edited Nearest Neighbours
ViT	Vision Transformers
XGBOOST	eXtreme Gradient Boosting

CHAPTER 1: INTRODUCTION

1.1 Background

1.1.1 Concept

Skin cancer, particularly melanoma, is one of the most aggressive forms of cancer, necessitating early detection for effective treatment and improved patient outcomes. Computer-aided diagnosis (CAD) systems using deep learning have shown great potential in automating skin lesion classification with high accuracy. The ISIC 2024 dataset, consisting of dermoscopic images and patient metadata, provides a valuable resource for training such models.

Deep learning models, especially Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), have revolutionized medical image analysis by enabling automated feature extraction and lesion classification. Additionally, multimodal learning, which integrates image and metadata, has proven to enhance diagnostic performance. However, challenges such as class imbalance, varying image quality, missing data, and complex feature interactions necessitate advanced preprocessing and model optimization techniques. This study leverages segmentation models (DeepLabv3+), CNN-based classifiers (ResNet, DenseNet), transformer architectures (ViT, MedMamba), and ensemble learning (XGBoost, LightGBM, CatBoost) to achieve accurate classification incorporating the multimodal input, while addressing these challenges.

1.1.2 Motivation

Despite the rapid progress in AI-driven medical diagnosis, skin lesion classification remains a challenging task due to the high variability in lesion appearance, dataset imbalances, and the need for clinically relevant sensitivity. Traditional diagnostic methods rely on derma-

tologists' expertise, which, while effective, is time-consuming and subject to inter-observer variability. Thus, there is a growing demand for AI-powered solutions that can assist clinicians in identifying malignant lesions with high sensitivity and specificity.

The primary motivation for this project is to develop a robust, scalable, and clinically reliable deep learning model that can effectively classify benign and malignant lesions. By leveraging state-of-the-art architectures and multimodal learning approaches, this research aims to improve upon existing models and contribute to the field of AI-assisted dermatology. Furthermore, the focus on handling class imbalance, refining lesion segmentation, and optimizing model architectures ensures that the developed system is both practical and adaptable for real-world clinical applications.

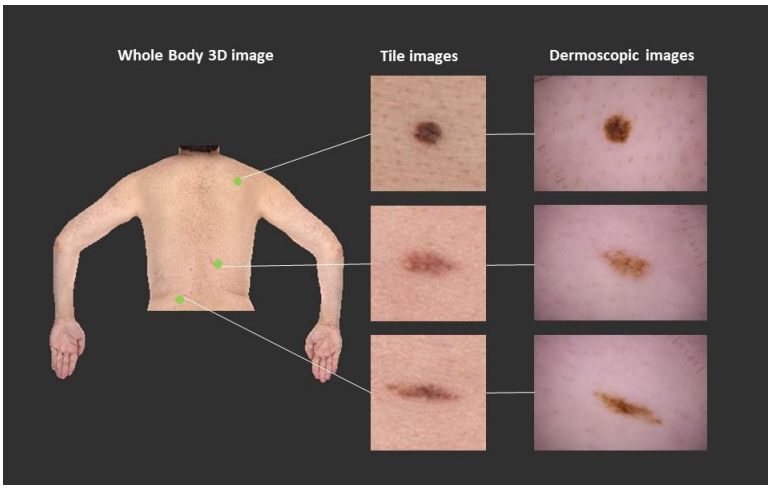

Figure 1.1: Sample Dermoscopy Images

CHAPTER 2: LITERATURE SURVEY

2.1 Introduction to Skin Lesion Analysis and the ISIC Challenges

The analysis of skin lesions, particularly towards melanoma detection, has been a significant area of research in biomedical imaging. The International Skin Imaging Collaboration (ISIC) has hosted challenges, such as the one in 2017, to advance this field. Furthermore, the ISIC 2019 competition focused on the classification of skin lesions into eight distinct classes, indicating the complexity and the need for robust classification methods. These challenges highlight the importance of developing effective techniques for accurate skin lesion analysis.

2.2 Deep Learning for Skin Lesion Classification

Deep learning techniques, especially deep convolutional neural networks (CNNs) and transfer learning, have been successfully applied to the task of skin lesion classification. The success of deep learning is attributed to its self-learning and generalization abilities, particularly when trained on large amounts of data.

2.3 Multimodal Medical Image Analysis

Multi-modality is a widely adopted approach in medical imaging as it can provide diverse information about the target, such as a tumor, organ, or tissue. Segmentation using multi-modality involves the fusion of information from different modalities to enhance the accuracy of segmentation. Recently, deep learning-based approaches have gained significant traction in multi-modal medical image segmentation due to their ability to learn complex patterns and generalize from large datasets. [1]

Various deep learning network architectures are employed for multi-modal segmentation, and their fusion strategies are crucial for performance. These strategies can be broadly categorized as earlier fusion and later fusion. Earlier fusion methods are simpler and integrate the information at an early stage, focusing on the subsequent segmentation network's architecture. Conversely, later fusion methods pay more attention to the fusion strategy itself to learn the intricate relationships between different modalities. Generally, if the fusion method is sufficiently effective, later fusion can yield more accurate results compared to earlier fusion. Common challenges exist in medical image segmentation, necessitating the development of robust techniques. [2]

2.4 Multimodal Self-Supervised Learning

Self-supervised learning offers a way to leverage unlabeled data to learn generic representations, which can then be used for more annotation-efficient learning on downstream tasks. A novel approach in this domain is multimodal self-supervised learning, which utilizes multiple imaging modalities to acquire this generic knowledge. One such method introduces the multimodal puzzle task, where representations are learned by confusing image modalities at the data level. The Sinkhorn operator is used to formulate the puzzle-solving optimization as permutation matrix inference, allowing for efficient handling of multimodal puzzles with varying complexity. Furthermore, generation techniques can be used for multimodal data augmentation during self-supervised pretraining. This aims to mitigate the quality issues associated with purely synthetic images and improve the data efficiency and the quality of the learned representations. Experimental results have shown that solving these multimodal puzzles leads to better semantic representations compared to processing each modality independently. The use of synthetic images for self-supervised pretraining has also demonstrated benefits. This approach has been showcased on segmentation tasks, achieving competitive results compared to state-of-the-art methods. [3]

2.5 Addressing Challenges in Data-Scarce Environments with GANs and Few-Shot Learning

Many practical fields, including drug discovery, medical health records, and malicious traffic detection, face the challenge of limited effective data due to confidentiality, scarcity, and high acquisition costs. Developing machine learning models that perform well with small amounts of data is therefore crucial for widespread application. Few-shot learning aims to tackle this problem and can be approached from three main perspectives: data, models, and optimization algorithms. [4]

Data-based few-shot learning methods focus on enhancing the dataset, either at the sample level (e.g., data augmentation) or at the feature level (e.g., feature selection and optimization). While intuitive, data augmentation can introduce noise or bias.

Model-based few-shot learning methods address the problem through model design, including model fine-tuning (leveraging knowledge from related domains) and metric learning (learning suitable metrics for small-sample data). These methods often require related datasets.

Optimization algorithm-based few-shot learning methods modify the optimal hypothesis search, with meta-learning being a prominent approach. Meta-learning trains a meta-learner on multiple tasks to enable quick learning of new tasks by modifying the optimization algorithm. [5]

Generative Adversarial Networks (GANs) learn data distributions and generate new data using maximum likelihood principles and adversarial learning. A GAN consists of a generator that aims to learn the data distribution and a discriminator that aims to distinguish between real and generated samples. Through adversarial training, the model learns to generate new samples that resemble the training data. However, in few-shot scenarios, the distributions learned by both the generator and discriminator can exhibit biases. [6]

To address these biases, techniques like reparameterization GANs combined with Markov Chain Monte Carlo (MCMC) sampling can be used to correct the generator's learned distribution. Additionally, ensemble methods can constrain discriminator learning and correct its learned distribution. One proposed architecture, DAMFT_FSL2, incorporates a reparameterization GAN ensemble into the data augmentation module and uses increased iteration rounds and MHLoss in the model fine-tuning module to enhance stability and classification performance. This involves an ensemble strategy for the discriminator and MCMC sampling for the generator to obtain more relevant datasets. The discriminator bias can be reduced using ensemble strategies like Bagging, which improves the discriminator's stability. Generator bias can be corrected using MCMC sampling, with the calibrated discriminator's implicit distribution serving as the target distribution. Fine-tuning strategies, such as increasing iteration rounds and using MHLoss, can improve the stability and convergence of model training on few-shot data. [7]

Experiments have shown that MCMC sampling and discriminative model ensemble strategies effectively enhance the realism of generated data, with their combination leading to further improvements. The MhERGAN algorithm[8], based on reparameterized GAN ensemble and MHLoss fine-tuning, has demonstrated superior performance in few-shot learning scenarios compared to other methods, highlighting the benefits of bias correction and fine-tuning stability improvements, particularly for small-sample data.

2.6 Gap Analysis

There is limited application and evaluation of multimodal data integration (e.g., combining dermoscopic images with clinical metadata) specifically for skin lesion classification. Multimodal self-supervised learning (e.g., multimodal puzzle tasks) has proven effective for segmentation tasks. These approaches have not been widely adopted or optimized for skin lesion classification datasets like ISIC, where labeled data is scarce. Few studies have explored cost-sensitive learning, balanced loss functions, or bias-corrected GANs specifically

for skin lesion datasets, which typically suffer from severe class imbalance. GAN-based and meta-learning few-shot methods show effectiveness across several domains. These methods have not been sufficiently adapted or evaluated in the context of dermatological imaging, where small, imbalanced, and noisy datasets are common. The potential of longitudinal data (e.g., follow-up images of lesions over time) remains largely unexplored. Few works have incorporated explainable AI (XAI) methods, and most are limited to unimodal CNNs. There's a lack of clinical validation of explanation methods in multimodal ensemble models. High-performing models often use large architectures that are computationally expensive. Lightweight models (e.g., MobileViT, efficient Transformer variants) have not been deeply investigated for mobile or point-of-care dermatological applications. So, we try to tackle the above mentioned limitations in our research study.

CHAPTER 3: METHODOLOGY

3.1 Data Collection and Preprocessing

3.1.1 ISIC 2024 Dataset

The primary dataset employed in this study is the ISIC 2024 Challenge Dataset, a comprehensive and large-scale collection curated for the advancement of skin lesion analysis using machine learning and computer vision techniques. This dataset comprises over 400,000 high-resolution RGB dermoscopic images, capturing a wide array of skin lesion types, including both benign and malignant cases. These images are designed to reflect real-world clinical variability, encompassing differences in lighting, skin tones, and lesion appearances, thereby simulating the complexities encountered in actual dermatological assessments.

In addition to the image data, the dataset is accompanied by extensive metadata, which includes patient demographics such as age and gender, as well as anatomical site information indicating the location of the lesion on the body. This multimodal nature of the dataset allows for more robust feature extraction and supports the development of models that can incorporate both visual and contextual information, potentially improving diagnostic accuracy.

One of the most critical challenges associated with this dataset is the severe class imbalance. Out of the entire dataset, only 393 samples are labeled as malignant, while the vast majority belong to the benign category. This imbalance poses a significant risk of model bias, where the learning algorithm may default to predicting the majority class, leading to high accuracy but poor sensitivity towards the malignant class. Consequently, it becomes essential to implement specialized strategies to handle this imbalance. These may include data augmentation techniques, re-sampling methods (e.g., oversampling the minority class

or undersampling the majority), cost-sensitive learning, and custom loss functions like focal loss or class-weighted cross-entropy, all of which aim to enhance the model's ability to detect and learn patterns associated with rare malignant lesions.

Moreover, addressing this imbalance is not only crucial for improving model performance but also has real-world clinical implications. In practice, failing to identify malignant lesions can have serious consequences, making sensitivity to the minority class a top priority in model development and evaluation.

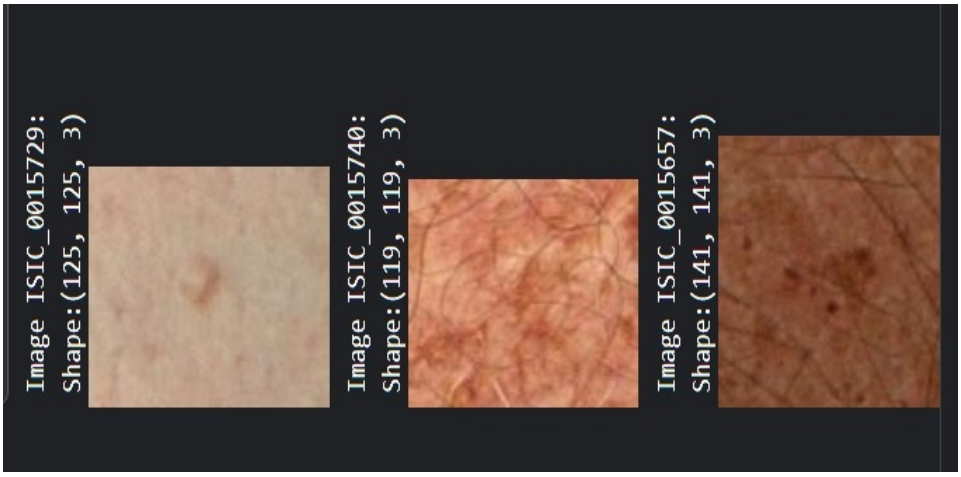

Figure 3.1: Test Samples

3.1.2 Handling Missing Data

Missing data is a prevalent issue in medical datasets and can significantly hinder the performance, generalizability, and trustworthiness of machine learning models. In the context of this study, several imputation strategies were adopted to address missing values across various features, ensuring data completeness and preserving the integrity of the dataset during model training and evaluation.

To handle complex, non-linear interactions in the data, Multiple Imputation by Machine Learning techniques were utilized. One such method is the Random Forest Imputer, which leverages the predictive power of ensemble learning. This imputer builds an ensemble of decision trees using the observed (non-missing) data and learns underlying patterns and feature correlations to predict the missing values. Being a non-parametric method, it does not

assume any specific data distribution, making it particularly suitable for medical datasets where relationships among variables can be highly non-linear and irregular. [9]

Another advanced technique used is the Iterative Imputer, which is based on a form of chained equations. This approach models each feature with missing values as a function of other features, typically using Bayesian Ridge Regression or other estimators. The process is repeated iteratively, where each round updates the imputed values and refines them further until convergence. This method ensures that imputations are statistically consistent and coherent, improving the quality of the reconstructed data matrix. [10]

The K-Nearest Neighbors (KNN) Imputer was also applied, especially in cases where the data was structured and had well-defined clusters [11]. This method computes the distance (typically Euclidean) between a data point with missing values and other complete samples, identifies the nearest neighbors, and imputes the missing value based on an average (for continuous features) or majority vote (for categorical features). This instance-based learning technique helps preserve local structures and patterns in the data.

For simpler or less critical features—particularly those with near-uniform distributions or minimal influence on model performance—basic imputation strategies were employed. These included filling missing values with the mean, median, or mode of the feature, depending on its distribution and type. These methods, while simplistic, are computationally efficient and effective when the missingness is random and infrequent.

Additionally, for features representing temporal sequences or ordered records, forward-fill and backward-fill techniques were used. These methods propagate the last known (or next known) valid value across missing entries in time-series-like data [12]. Such techniques are especially useful in scenarios where values are expected to change slowly over time or remain consistent within short intervals.

By combining multiple imputation methods tailored to different data types and contexts, this approach ensured a more robust and comprehensive handling of missing data, which is

crucial for the reliability of downstream machine learning models in medical applications.

3.1.3 Image Preprocessing

Prior to segmentation and model training, the dermoscopic images underwent a comprehensive preprocessing pipeline designed to enhance image quality, standardize input data, and optimize model learning efficiency. Given the diversity and variability inherent in clinical dermoscopy images—such as differing resolutions, lighting conditions, and artifacts—this step was critical to ensure both robustness and generalizability of the downstream machine learning models [13].

First, all images were resized to multiple resolutions, specifically 72×72, 128×128, and 224×224 pixels, to accommodate various deep learning architectures. Lower resolutions such as 72×72 were useful for lightweight or preliminary models and faster experimentation, while higher resolutions like 224×224 were ideal for models based on deep convolutional neural networks (CNNs) and Vision Transformers (ViTs) that benefit from finer spatial details. This multi-resolution approach also facilitated model comparison and ensemble strategies.

To ensure consistent intensity ranges across the dataset, pixel value normalization was applied. All pixel intensities were scaled to a [0, 1] range by dividing by 255, standardizing the input distribution and helping to accelerate convergence during training by stabilizing gradient updates. In some cases, z-score normalization or per-channel mean subtraction was also considered, depending on the model requirements.

To further increase the diversity and quantity of training data, data augmentation techniques were applied extensively. These transformations simulated various real-world conditions and imaging inconsistencies to help models generalize better [14]. Augmentations included:

- Horizontal and vertical flipping to introduce spatial variability,

- Random rotations to account for orientation differences in lesion capture,
- Brightness and contrast adjustments to simulate lighting variations,
- Gaussian noise addition to improve model robustness to pixel-level noise and sensor artifacts.

By implementing these preprocessing techniques, the dataset was transformed into a cleaner, more uniform, and information-rich format suitable for deep learning. This stage was instrumental in enhancing model performance, minimizing overfitting, and ensuring that the key diagnostic features of skin lesions were preserved and emphasized.

3.2 Image Segmentation

3.2.1 DeepLabv3+ for Lesion Segmentation

Accurate lesion segmentation is a foundational step in automated skin lesion analysis, as it enables the model to focus on the region of interest (ROI) while excluding irrelevant background, surrounding skin, and common artifacts such as hair or rulers. Effective segmentation not only improves classification accuracy but also enhances the interpretability of the model by localizing pathological regions.

In this study, the DeepLabv3+ architecture was adopted for the task of semantic segmentation, owing to its state-of-the-art performance in a wide range of medical image analysis applications. DeepLabv3+ extends the original DeepLabv3 model by integrating a decoder module, which significantly improves boundary delineation and spatial localization—critical requirements in dermoscopic image segmentation. [15]

At the core of DeepLabv3+ is the Atrous Spatial Pyramid Pooling (ASPP) module. ASPP captures rich, multi-scale contextual information by applying dilated (atrous) convolutions with varying dilation rates in parallel. This allows the model to analyze the lesion at multiple scales simultaneously, which is especially important for skin lesions that vary dramatically

in terms of size, shape, and texture. ASPP enhances the model's ability to detect both fine-grained details and broader structural patterns within the lesion.

The encoder of DeepLabv3+ is built using the Xception backbone—a deep convolutional neural network that employs depthwise separable convolutions to reduce computational cost while maintaining high representational power. This backbone efficiently extracts high-level semantic features from input images while preserving spatial resolution, enabling the model to retain important boundary and texture information.

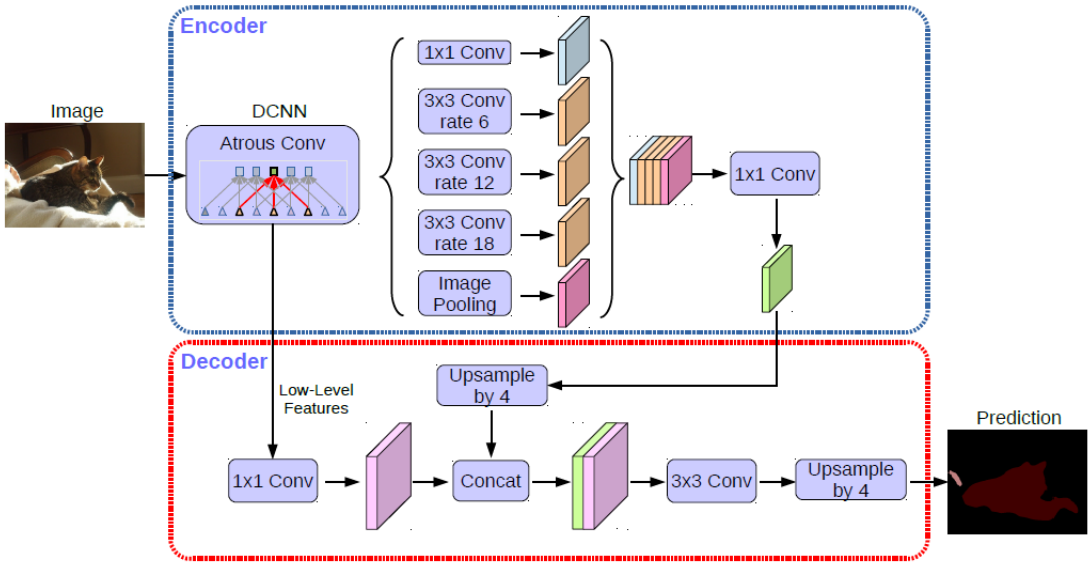

Figure 3.2: DeepLabv3+ Architecture

The decoder module further refines the segmentation output by combining high-level semantic features from the encoder with lower-level spatial features. It uses bilinear upsampling and convolutional refinement layers to reconstruct precise lesion boundaries. This is crucial in medical contexts, where even slight inaccuracies in segmentation can affect downstream diagnostic decisions.

To post-process and further enhance the quality of the segmented masks, morphological operations were employed. Specifically, erosion was used to eliminate small, isolated pixels or noise near the lesion edges, while dilation helped restore the overall shape of the lesion by expanding valid regions. These operations contributed to cleaner and more anatomically accurate segmentations.

In addition, Conditional Random Fields (CRFs) were applied as a final refinement step. CRFs are probabilistic graphical models that enforce spatial consistency and improve boundary sharpness by modeling relationships between neighboring pixels. This technique helps in aligning the predicted segmentation with the true lesion boundary, especially in cases where lesions are indistinct or surrounded by low-contrast skin.

Together, this segmentation pipeline—comprising DeepLabv3+ with ASPP and Xception, followed by morphological operations and CRF-based refinement—ensures a highly accurate, smooth, and clinically useful delineation of skin lesions, setting a strong foundation for subsequent tasks such as classification or malignancy prediction.

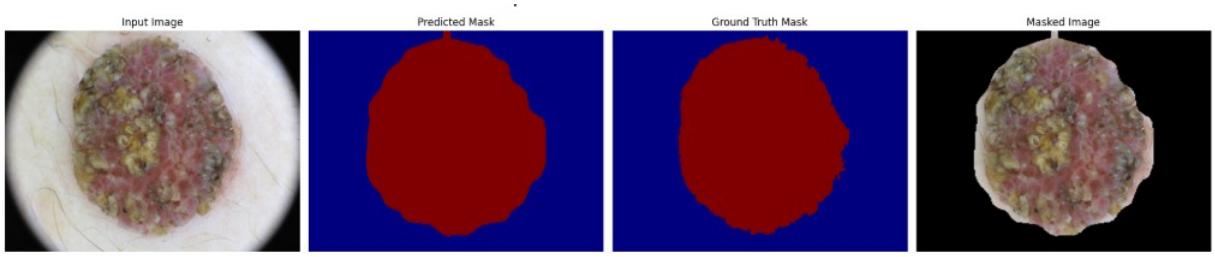

Figure 3.3: Segmented image samples

3.3 Feature Selection and Engineering

3.3.1 Feature Transformation and Engineering

While image data forms the core of dermoscopic lesion analysis, metadata—such as patient age, sex, and anatomical site—provides crucial contextual information that can significantly enhance model performance when appropriately processed. However, raw metadata often requires careful transformation to become compatible with machine learning models, especially when combining tabular and visual modalities.

The first step involved cleaning and filtering the metadata. Columns deemed irrelevant to prediction—such as lesion identifiers, image filenames, or acquisition-specific tags—were removed to prevent data leakage and reduce unnecessary dimensionality. These fields could

inadvertently allow the model to memorize specific instances rather than learning generalizable patterns.

Next, categorical variables such as sex (e.g., male, female, unknown) and anatomical site (e.g., back, upper extremity, scalp) were transformed using one-hot encoding. This process converts each category into separate binary features (0 or 1), allowing models to interpret them numerically without assuming any ordinal relationship. While one-hot encoding is effective for low-cardinality variables, it can lead to high sparsity when applied to features with many unique values.

To address this, for high-cardinality categorical features, feature hashing (also known as the hashing trick) was used. This technique maps categories to a fixed-size numerical vector using hash functions, thereby reducing memory usage and dimensionality while still capturing distinguishing characteristics. Although feature collisions can occur, in practice, hashing provides a good trade-off between complexity and representation power—especially in large-scale datasets like ISIC.

For continuous numerical features such as age, two types of normalization techniques were applied based on the data distribution:

- MinMaxScaler normalized values into a fixed [0, 1] range, useful for features with known bounded ranges.
- StandardScaler standardized features to have zero mean and unit variance, helping models that are sensitive to feature scale (e.g., gradient-based optimizers in neural networks).

To further optimize the feature space and eliminate redundancy, dimensionality reduction was performed using Principal Component Analysis (PCA). PCA transforms the original features into a new orthogonal basis, where each principal component captures the maximum variance. This process retains the most informative aspects of the data while removing noise and collinear relationships. It also helps in visualizing high-dimensional metadata in lower-

dimensional subspaces (e.g., 2D or 3D), aiding exploratory data analysis.

In addition to unsupervised techniques like PCA, feature selection was guided by importance scores from tree-based models such as XGBoost and Random Forests. These models naturally compute feature importance during training, highlighting variables that contribute most to classification decisions. By ranking and selecting the top features based on their importance scores, the model pipeline was streamlined to focus only on the most predictive and relevant variables, leading to improved training efficiency and potentially better generalization.

Together, these preprocessing steps transformed raw metadata into a well-structured, machine-readable format, allowing it to be seamlessly integrated with visual data inputs or used independently in hybrid or ensemble models.

3.4 Handling Class Imbalance

Class imbalance posed a significant challenge in this project, given the disproportionate distribution between benign and malignant skin lesion samples in the ISIC 2024 dataset. Out of over 400,000 dermoscopic images, only 393 were labeled as malignant, highlighting the extreme rarity of the positive class. This imbalance can severely bias learning algorithms, causing them to favor the majority (benign) class and underperform on the minority (malignant) class—which is of highest clinical significance.

To address this, a multi-pronged approach was implemented, combining data-level resampling strategies with algorithm-level modifications to improve the model's sensitivity to malignant cases while minimizing overfitting or bias. [16]

1. Random Undersampling

   As a preliminary balancing step, random undersampling was used to reduce the number of benign samples to a more manageable size. While this brought class proportions closer together and reduced training time, it came at the cost of discarding potentially

valuable benign examples, risking the loss of important variability and lesion diversity.

2. Synthetic Oversampling with SMOTE

   To counteract the rarity of malignant cases, SMOTE (Synthetic Minority Over-sampling Technique) was employed. SMOTE works by generating synthetic malignant examples through interpolation between neighboring malignant samples in the feature space. This helps to create a smoother decision boundary and alleviate the bias toward the majority class. However, SMOTE can also inadvertently generate ambiguous or noisy samples, particularly near class boundaries.

3. Hybrid Sampling Techniques: SMOTETomek and SMOTEENN

   To refine synthetic sample generation and improve data quality, hybrid methods were introduced:

   - SMOTETomek combines SMOTE with Tomek links, which identify and remove borderline samples where a sample from one class is the nearest neighbor of a sample from the other class. This helps to clean noisy or overlapping regions in the feature space, improving the clarity of class boundaries.

   - SMOTEENN (Edited Nearest Neighbors) integrates SMOTE with an instance cleaning technique that removes samples misclassified by their k-nearest neighbors. This not only smooths class boundaries but also eliminates mislabeled or ambiguous samples, further reducing noise introduced by oversampling.

4. NearMiss for Informed Undersampling

   In addition to random undersampling, NearMiss algorithms were applied to intelligently select benign samples that are most similar to malignant cases based on distance metrics. By focusing on hard negatives—benign lesions that resemble malignant ones—the model is encouraged to learn fine-grained distinctions, which enhances its ability to differentiate between visually similar lesions.

5. Class Weighting in Loss Functions

   Beyond data augmentation, algorithm-level balancing was enforced by incorporating class weights directly into the model's loss function. This ensures that misclassifying a malignant sample incurs a higher penalty than misclassifying a benign one. For neural network-based models, this was implemented by modifying the categorical cross-entropy loss with inverse frequency weighting or using focal loss, which further concentrates the learning on hard-to-classify samples. This technique is particularly effective when data augmentation alone is insufficient to overcome extreme imbalance.

By combining resampling, hybrid cleaning techniques, and loss function adjustment, the imbalance problem was tackled from multiple angles. This holistic strategy improved the model's ability to detect malignant lesions with higher sensitivity and precision, ensuring that the minority class was properly represented during training and evaluation without compromising overall generalization.

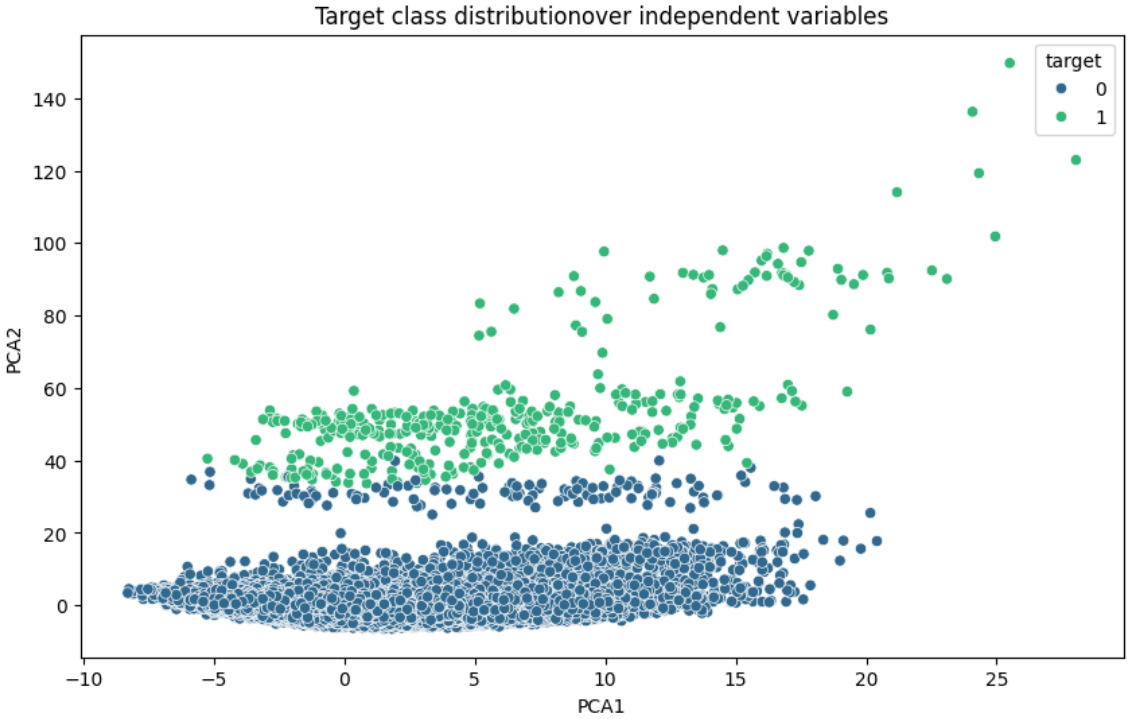

Figure 3.4: Target class distribution

3.5 Image Feature Extraction

3.5.1 CNN-Based Methods

To perform automated diagnosis from dermoscopic images, Convolutional Neural Networks (CNNs) were employed due to their proven capability in learning spatial hierarchies and extracting discriminative features from visual data. CNNs have become the cornerstone of modern computer vision tasks, particularly in medical imaging, where they can learn complex patterns directly from pixel data, eliminating the need for manual feature engineering.

Our methodology incorporates both custom-designed Deep Convolutional Neural Networks (DCNNs) and Transfer Learning techniques utilizing pre-trained models to achieve robust and efficient performance. Transfer learning not only accelerates convergence during training but also improves generalization, especially when working with relatively small or imbalanced datasets. ResNet and DenseNet were fine-tuned, each offering unique architectural advantages. [17]

We utilized the following pre-trained models from the kerashub models:

- ResNet Family: ResNet-18, ResNet-50, ResNet-101, ResNet-152, ResNetV2-50

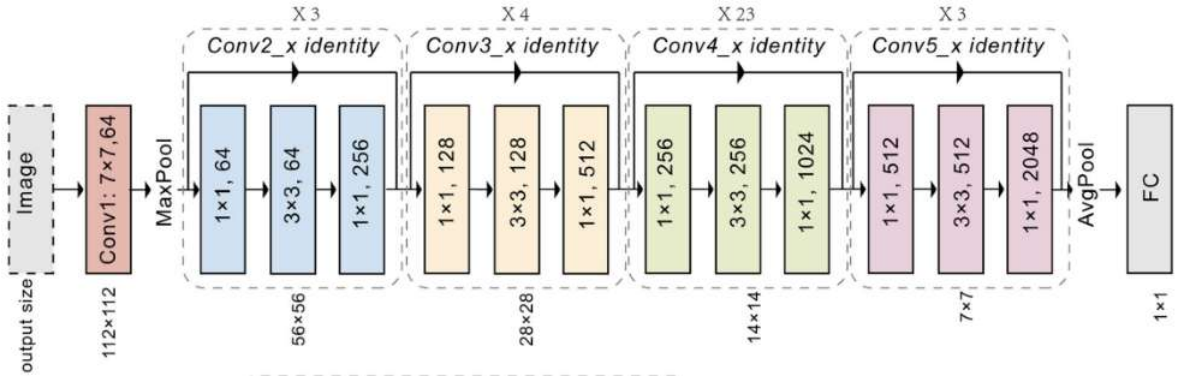

Figure 3.5: Resnet Architecture

- DenseNet Family: DenseNet-121, DenseNet-169, DenseNet-201

The general methodology for utilizing these pre-trained models involved the following

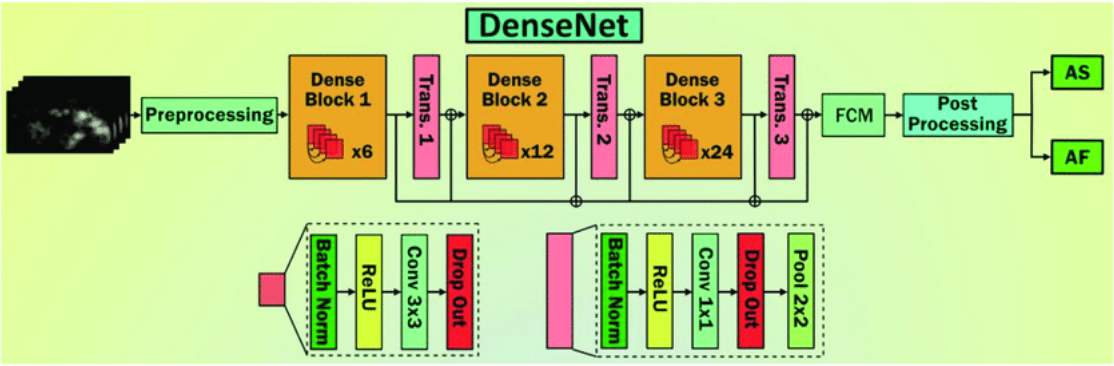

Figure 3.6: Densenet Architecture

steps:

- Loading Pre-trained Weights: We loaded the pre-trained weights of the chosen models, which were trained on the massive ImageNet dataset for image classification. These weights encapsulate a vast amount of knowledge about general visual features.
- Feature Extraction: We removed the final classification layer (fully connected layer with softmax) of the pre-trained model. The remaining network served as a powerful feature extractor, transforming our input images into high-dimensional feature vectors.
- Global Average Pooling (GAP): To further reduce the number of trainable parameters and obtain a fixed-size feature vector regardless of the input image size, we applied Global Average Pooling (GAP) to the output feature maps of the last convolutional layer of the pre-trained model. GAP computes the average of each feature map, resulting in a compact feature representation.
- Fine-tuning : In some experiments, we also explored fine-tuning the pre-trained weights. This involved unfreezing some or all of the layers of the pre-trained model and continuing to train them on our specific dataset with a smaller learning rate. Fine-tuning can further adapt the learned features to our specific task but requires careful consideration to avoid overfitting, especially with limited data.

To stabilize training and accelerate convergence, Batch Normalization was applied after

convolutional layers. This technique normalizes activations to maintain consistent distributions across layers, helping the model learn faster and generalize better by reducing internal covariate shift. Additionally, Dropout was used as a regularization method to prevent co-adaptation of neurons. By randomly deactivating a subset of neurons during training, dropout forces the network to learn more robust and redundant representations, thereby reducing the risk of overfitting.

These models were trained using standard optimization techniques such as Adam or SGD with appropriate learning rate schedules, and their performance was evaluated using metrics like accuracy, AUC-ROC, precision, recall, and F1-score—especially focusing on the malignant class, given its critical importance in clinical diagnosis.

By integrating these CNN architectures with careful tuning and regularization techniques, the models demonstrated strong ability to differentiate between benign and malignant lesions, marking an essential step towards AI-assisted dermatological screening.

3.5.2 Transformer-Based Methods

To complement the spatial feature extraction capabilities of Convolutional Neural Networks (CNNs), transformer-based models were integrated into the pipeline to leverage their superior ability to model long-range dependencies and capture global contextual information. Unlike CNNs, which are inherently local in their receptive fields, transformer models use self-attention mechanisms that enable each part of the image to dynamically attend to all other parts, making them particularly powerful for complex medical image analysis where fine-grained context matters. [18]

One of the core models employed was the Vision Transformer (ViT). ViT treats an input image as a sequence of flattened patches (e.g., 16×16), which are linearly embedded and passed through a position encoding layer followed by multiple transformer encoder blocks. Each block applies multi-head self-attention and feed-forward layers to model both local and global features.

To address ViT's reliance on large datasets for training, the Data-efficient Image Transformer (DeiT) was used. DeiT enhances ViT through a knowledge distillation approach, where a powerful CNN "teacher" model guides the training of a more compact transformer "student" model. This allows DeiT to achieve competitive performance with less data and lower computational cost, making it suitable for medical datasets that often suffer from limited sample sizes and class imbalance. The use of a distillation token in DeiT enables it to jointly learn from both ground-truth labels and the soft predictions of the teacher, effectively compressing knowledge into the transformer architecture. [19]

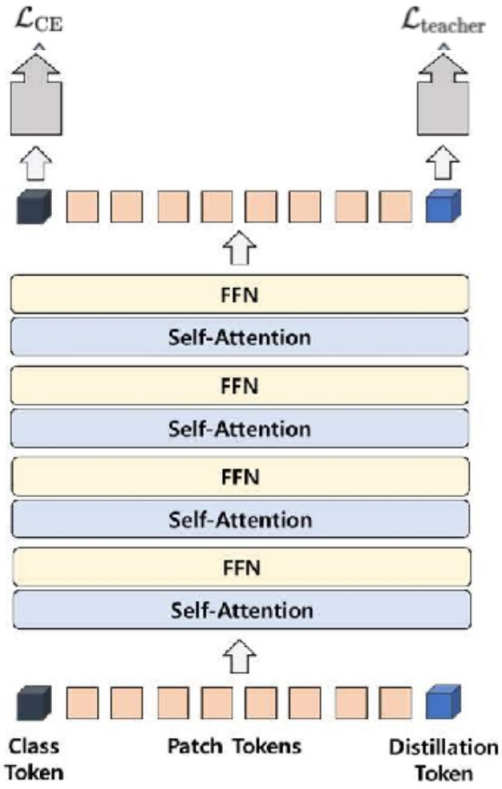

Figure 3.7: DeiT Architecture

Further extending transformer capabilities to mobile and edge applications, MobileViT was explored as a hybrid model that combines the strength of CNNs for local feature extraction with transformers for capturing global context. MobileViT integrates convolutional

layers at the initial stages to detect fine-grained patterns and transformer blocks in later stages to understand broader spatial relationships. This design allows MobileViT to retain a compact footprint while still delivering competitive accuracy, making it ideal for deployment in real-time or resource-constrained clinical environments.

To specialize transformer architectures for medical imaging, the study incorporated MedMamba, a domain-tailored variant inspired by the recently introduced Mamba architecture. MedMamba adapts Mamba's core strengths—efficient sequence modeling and long-range dependency learning—to 2D medical image data. Its architecture starts with a patch embedding layer, which converts the input image into smaller patches. These patches are then processed through a sequence of SS-Conv-SSM blocks, where SS-Conv refers to spatially selective convolution layers, and SSM (Selective State Space Models) extends Mamba's 1D design into 2D. The blocks are interleaved with patch merging layers to perform hierarchical downsampling and progressively build high-level representations. [20]

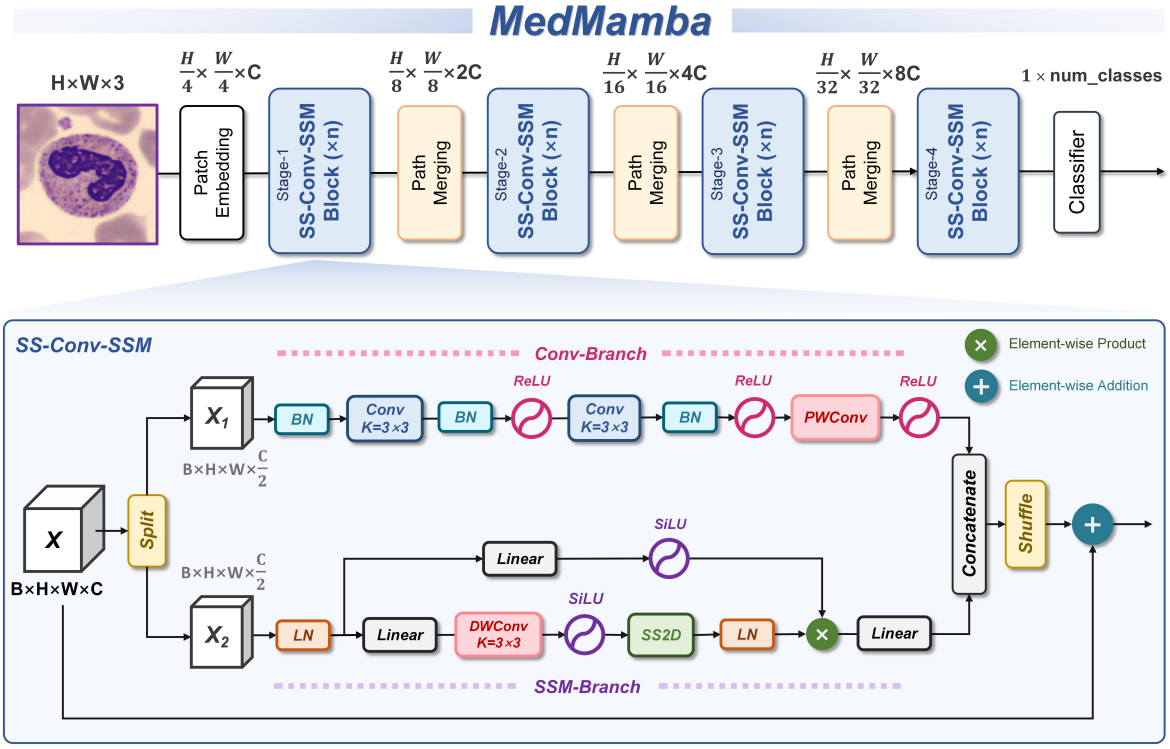

Figure 3.8: Medmamba Architecture

One of MedMamba's key innovations is the 2D-Selective-Scan (SS2D) module, inspired

by VMamba, which generalizes the concept of selective scanning from 1D sequences to 2D images. The Cross-Scan Module (CSM) implements a four-way scanning mechanism—left-to-right, right-to-left, top-down, and bottom-up—enabling the model to aggregate information across all spatial directions efficiently. This directional awareness is especially useful in medical imaging, where lesions may have irregular shapes and orientations.

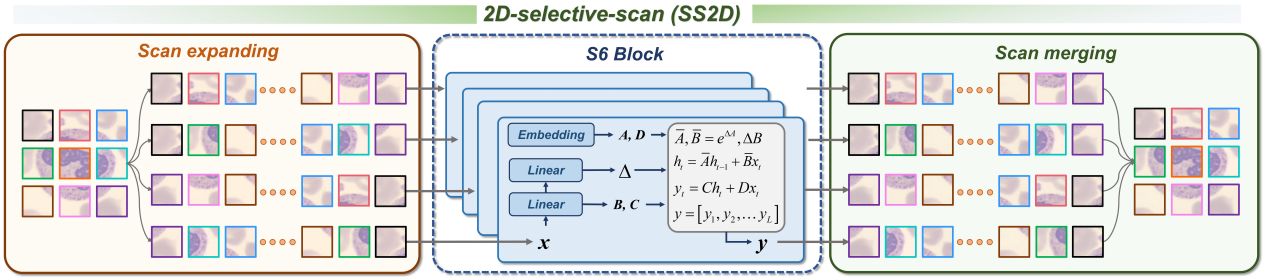

Figure 3.9: SS2D Architecture

MedMamba also emphasizes interpretability and diagnostic relevance, incorporating multi-scale attention mechanisms to highlight areas of medical importance while maintaining computational efficiency. Its domain-aware design enables it to outperform generic transformers on tasks such as lesion classification, where subtle visual cues are critical for distinguishing malignant features.

Together, these transformer architectures—ViT, DeiT, MobileViT, and MedMamba—offered a diverse set of tools for modeling both local and global aspects of skin lesion images. Their integration provided significant improvements in classification accuracy, robustness to noise, and the ability to generalize across different lesion types and image acquisition settings.

3.6 Classification Models

3.6.1 Deep Learning Classifiers

The classification stage was meticulously designed to integrate and leverage the complementary strengths of both tabular patient metadata and deep visual features extracted from

dermoscopic images. This was achieved through a carefully constructed feature fusion strategy that allowed the model to reason with multimodal information in a unified manner. The primary motivation for this fusion approach was to capture both the clinical context—such as patient demographics and lesion characteristics—and the intricate visual patterns seen in malignant skin lesions, such as asymmetry, irregular borders, and abnormal pigmentation. By combining these sources, the model could make more informed predictions, improving its sensitivity to subtle malignancy cues that might be missed when using either modality in isolation.

1. Deep Feature Extraction from Images

   On the imaging side, dermoscopic lesion images were passed through powerful deep learning models like DenseNet-169, Vision Transformer (ViT), and MedMamba. These architectures were selected for their ability to extract complex and hierarchical features from medical images. DenseNet-169 provided dense connections across layers, enabling feature reuse and efficient gradient flow. Vision Transformer (ViT) brought in the power of self-attention, which allowed the model to capture global context and long-range dependencies across the image. MedMamba, a recent addition to the transformer family tailored for medical image analysis, combined convolutional and state-space modeling to enhance the spatial coherence of learned features.

   These models transformed the raw image into a dense, high-dimensional embedding—essentially a compact vector of learned features that preserved critical visual information such as lesion structure, texture gradients, and color transitions. These embeddings served as a semantic representation of the image and were later used in conjunction with metadata for classification.

2. Metadata Transformation and Feature Preparation

   In parallel, structured patient metadata—containing attributes like age, sex, lesion location, and lesion history—underwent a rigorous preprocessing pipeline. As detailed

in previous sections, this involved normalization, encoding of categorical variables, and possibly dimensionality reduction. The result was a clean and standardized numerical feature vector that effectively represented the clinical side of the diagnosis.

To prepare for fusion, the metadata features and image embeddings were each passed through separate dense (fully connected) layers. These intermediate layers served a dual purpose: first, they allowed the model to learn modality-specific representations and patterns; second, they aligned the dimensionality of the two vectors, enabling smooth concatenation and ensuring that neither modality dominated the fused representation.

3. Feature Fusion and Multilayer Classification

   Once transformed, the metadata vector and the deep image embedding were concatenated to form a single fused feature vector. This vector encapsulated both the visual intricacies of the lesion and the contextual clinical knowledge, providing a rich and diverse set of signals to the downstream classifier. The fusion enabled the model to learn complex interactions between the two domains—such as how certain lesion shapes might be more suspicious in older individuals or in specific anatomical regions.

   The fused vector was then passed into a deep neural network classifier, composed of multiple fully connected layers. LeakyReLU activation functions were used to introduce nonlinearity while avoiding the "dying ReLU" problem. Batch normalization layers stabilized learning by normalizing activations during training, and dropout was used as a regularization strategy to reduce overfitting and enhance generalization. These components collectively allowed the network to learn intricate decision boundaries that would be difficult to capture with simpler models.

4. Final Output and Interpretability

   The final layer of the classifier was a single neuron with a sigmoid activation function. This produced a value between 0 and 1, representing the model's confidence in

the lesion being malignant. This probability-based output allowed for threshold tuning based on clinical priorities—for instance, lowering the threshold to reduce false negatives in high-risk diagnostic scenarios.

By adopting this fusion-based approach to classification, the model was able to exploit multimodal synergy. Image-based models alone may overlook contextual clues like patient demographics, while metadata-only models may miss visual cues that signal malignancy. When combined, these modalities offer a powerful framework that mimics the multi-perspective reasoning process used by dermatologists. This significantly enhanced the model's robustness, interpretability, and diagnostic precision—critical requirements for deployment in clinical settings where accurate detection of malignant lesions can have life-saving implications.

3.6.2 Tree-Based Models

In addition to image-based deep learning models, the structured metadata accompanying the dermoscopic images—such as patient age, sex, and anatomical site—was leveraged using powerful tree-based ensemble models. These models are well-suited for tabular data and offer robustness to feature scaling, collinearity, and non-linear relationships, making them ideal for medical metadata classification tasks. [21]

1. XGBoost (Extreme Gradient Boosting)

   XGBoost was one of the primary models implemented due to its high efficiency and predictive accuracy. It builds gradient-boosted decision trees, where each new tree attempts to correct the errors made by the previous ones, optimizing a regularized loss function.

   - The inclusion of L1 (Lasso) and L2 (Ridge) regularization helped control model complexity and prevent overfitting.
   - XGBoost also offered advanced features such as tree pruning, parallelized exe-

cution, and handling of missing values internally, making it particularly effective for real-world medical datasets with occasional missing metadata.

2. LightGBM (Light Gradient Boosting Machine)

   LightGBM was used alongside XGBoost for its speed and scalability, especially on large datasets. It uses:

   - A histogram-based decision tree algorithm, which discretizes continuous features into buckets, speeding up computation and reducing memory usage.
   - Leaf-wise tree growth, which selects the leaf with the highest loss reduction to split, leading to deeper, more accurate trees compared to level-wise growth. These properties made LightGBM an attractive option for quick prototyping and tuning across various feature sets.

3. CatBoost

   CatBoost was incorporated for its native support of categorical variables, which are common in medical metadata (e.g., gender, lesion location). It internally performs target-based encoding while preventing target leakage through ordered boosting—a form of permutation-driven learning that ensures statistical soundness.

   - CatBoost eliminates the need for extensive preprocessing of categorical features, reducing the potential for human-introduced bias and error.
   - It is also robust to overfitting and handles class imbalance relatively well through its built-in loss function adjustments.

4. Ensemble Learning and Meta-Modeling

   While each model performed well independently, their strengths were complementary, and combining them improved both generalization and prediction stability:

   - A stacked ensemble was constructed, where the predictions (probability out-

puts) of XGBoost, LightGBM, and CatBoost were used as inputs to a meta-classifier—specifically, a logistic regression model. This allowed the ensemble to learn the optimal weighting of each base learner's output based on performance trends.

- Additionally, soft voting was applied, where the final probability predictions from each model were averaged, and the class with the highest mean probability was selected. Soft voting ensures smoother decision boundaries compared to hard voting, which only considers the majority class labels.

This ensemble approach combined the accuracy and interpretability of gradient boosting models with the flexibility of probabilistic fusion, producing a robust classifier that could complement the image-based deep learning models. By integrating predictions from both modalities—image and metadata—the overall diagnostic system achieved improved sensitivity, precision, and reliability.

3.7 Training and Optimization Techniques

3.7.1 Optimizers

Choosing the right optimizer is crucial for ensuring stable convergence, efficient learning, and high generalization in deep learning models. In this project, we utilized several adaptive optimization algorithms that dynamically adjust learning rates based on gradient history, enabling more refined and context-sensitive updates to the model parameters [22]. These optimizers are especially beneficial in training complex architectures such as CNNs, Vision Transformers, and hybrid models like MedMamba, where learning dynamics can vary widely across layers and tasks. Below, we discuss the optimizers used in detail:

1. Adam (Adaptive Moment Estimation)

   Adam is a widely adopted optimizer in deep learning that combines the benefits of mo-

mentum and adaptive learning rates. It computes exponentially decaying averages of past gradients and their squares to estimate the first and second moments, respectively. These estimates are then bias-corrected to ensure stability during early training iterations. Adam adapts the learning rate for each parameter individually, enabling faster convergence and improved performance in problems with noisy or sparse gradients. Its ability to handle non-stationary objectives and scale well across different architectures makes it a go-to choice for many neural network applications.

2. AdamW (Adam with Decoupled Weight Decay)

   AdamW builds upon the Adam optimizer by introducing a key improvement: the decoupling of weight decay from gradient updates. In traditional Adam, weight decay is implemented as L2 regularization, which can interact unfavorably with adaptive learning rates. AdamW resolves this by applying weight decay independently, leading to better regularization and reduced overfitting, especially in transformer-based models. This decoupled approach ensures that the weight decay directly controls the magnitude of weights without interfering with the gradient computation, thereby enhancing generalization without compromising convergence speed.

3. Nadam (Nesterov-accelerated Adaptive Moment Estimation)

   Nadam merges the concepts of Adam with Nesterov Accelerated Gradient (NAG). Unlike traditional momentum, where the update is computed from the current gradient, NAG computes the gradient at the look-ahead position—after applying the momentum term—resulting in smoother and more informed updates. Nadam leverages this by incorporating Nesterov momentum into Adam's adaptive framework, leading to faster convergence, especially in tasks where gradients exhibit high variance. It is particularly effective in fine-tuning deep networks where stability and responsiveness are both critical.

4. RMSprop (Root Mean Square Propagation)

RMSprop is another adaptive optimizer that adjusts the learning rate for each parameter based on the magnitude of its recent gradients. By maintaining an exponentially decaying average of squared gradients, RMSprop normalizes the updates, which prevents the learning rate from becoming too large in directions with steep curvature. This makes it especially effective in handling non-stationary objectives and recurrent neural networks (RNNs). In this project, RMSprop contributed to stability in scenarios where the dataset was imbalanced or where the gradient signals were noisy due to data augmentation or synthetic sample generation.

3.7.2 Loss Functions

The choice of loss function is central to model training, especially in binary classification tasks with severe class imbalance—as is the case in skin lesion diagnosis. A well-designed loss function not only guides the model to minimize prediction error but also influences its focus during learning, such as prioritizing minority classes or promoting confident predictions. In this project, a combination of loss functions was employed to strike a balance between sensitivity to rare malignant cases, robust convergence, and generalization capability [23]. The following are the primary loss functions used and their roles in the training pipeline:

1. Binary Cross-Entropy (BCE)

   Binary Cross-Entropy is the fundamental loss function for binary classification problems. It calculates the difference between the predicted probability and the actual binary class label using the formula:

$$\mathcal{L}_{\text{BCE}} = -\left[y \cdot \log(p) + (1-y) \cdot \log(1-p)\right]$$

   where y is the ground truth label and p is the predicted probability. This loss penalizes predictions based on their confidence and correctness, encouraging the model to assign high probabilities to true positives and low probabilities to true negatives. BCE is

sensitive to probabilistic output quality, making it a suitable baseline loss for evaluating model calibration and binary classification performance.

2. Focal Loss

Focal Loss extends BCE by focusing more on hard examples and less on well-classified ones. It is particularly effective in imbalanced classification tasks. The focal loss is defined as:

$$\mathcal{L}_{\text{Focal}} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t)$$

where:

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{if } y = 0 \end{cases}$$

α_t is the class weighting factor,

γ is the focusing parameter, typically set to values like 2.

This formulation reduces the loss contribution from easy samples and amplifies it for misclassified or difficult examples, helping the model focus more on the minority class—in this case, malignant lesions.[40]

3. Squared Hinge Loss

Squared Hinge Loss, originally derived from Support Vector Machines (SVMs), introduces a margin-based penalty that encourages the model to output not just correct predictions but also predictions with high confidence. It is defined as:

$$\mathcal{L}_{\text{Hinge}} = [\max(0, 1 - y \cdot \hat{y})]^2$$

where:

$y \in \{-1, +1\}$ is the true class label (note the SVM-style encoding),

$\hat{y}$ is the raw prediction (logit output from the model).

This loss increases rapidly when the model's prediction lies close to the decision boundary, thus pushing predictions further away from the margin. In this project, Squared Hinge Loss was applied in scenarios where higher confidence margins were critical, contributing to improved discriminative ability and robustness against minor input noise. It penalizes predictions that lie too close to the decision boundary, driving the model to output scores with a larger margin and thereby improving classification confidence and generalization.

3.7.3 Training Strategies

Efficient training of deep learning models, particularly in sensitive domains like medical imaging, requires more than just powerful architectures and high-quality data. Optimization strategies and training heuristics play a vital role in accelerating convergence, preventing overfitting, and ensuring the robustness of the resulting models. In this project, several complementary training techniques were integrated to enhance learning performance and stability. These include early stopping, dynamic learning rate scheduling, mini-batch training, transfer learning, and stratified data splitting—all of which contributed to the effective training of both image-based and metadata-based models.

1. Early Stopping

   Early stopping is a regularization technique that halts training when performance on a validation set stops improving. In this project, the validation loss was continuously monitored after each epoch. If the validation loss did not decrease for a predefined number of consecutive epochs (patience), the training was terminated. This strategy

helped avoid overfitting by stopping the training process before the model began to memorize noise or irrelevant patterns from the training data. It also reduced computational costs by eliminating unnecessary epochs once optimal performance was reached.

2. Learning Rate Schedulers

   In the context of our project on skin lesion malignancy classification, learning rate schedulers played a pivotal role in controlling the pace of learning during model training. Given the architectural depth of the models involved—ranging from convolutional neural networks (CNNs) like DenseNet-169 to transformer-based frameworks such as ViT and MedMamba—static learning rates were insufficient for achieving optimal convergence. Furthermore, the imbalanced nature of the ISIC 2024 dataset added an additional layer of complexity, necessitating dynamic and context-sensitive learning rate adjustment to avoid overfitting or suboptimal training plateaus. Following choice of learning rate schedulers were used in the model training,

   - PiecewiseConstant Scheduler : The learning rate was held constant within specific epoch ranges and then sharply dropped (e.g., after 10, 20, 30 epochs) to allow for finer adjustments as training progressed.Help stabilize training by reducing the learning rate at fixed epoch milestones. Beneficial for large models like ViT, ensuring aggressive learning early on and refined tuning later.

   - ExponentialDecay Scheduler: gradually reduce the learning rate after every step/epoch based on a decay rate, ensuring smooth convergence and reducing the risk of overshooting.

   - ReduceLROnPlateau Scheduler: Dynamically reduced the learning rate when validation loss plateaued. Monitored val loss and reduced LR if no improvement after patience epochs. Prevented overfitting and ensured continued learning progress when stuck in flat regions of the loss curve—especially effective for

models that tend to converge early or irregularly

3. Mini-Batch Training

   Mini-batch training processes the dataset in smaller subsets (batches) rather than feeding the entire dataset at once. This strategy strikes a balance between stochastic gradient descent (one sample at a time) and batch gradient descent (all samples at once). In this project, mini-batches enabled efficient use of computational resources (especially GPU memory), improved generalization by introducing gradient noise, and accelerated convergence. Additionally, batch normalization layers performed more effectively with mini-batches, further stabilizing the training process.

4. Transfer Learning

   Transfer learning leverages knowledge from pre-trained models—usually trained on large-scale datasets like ImageNet—to jumpstart the training process on a new but related task. In this project, architectures such as DenseNet-169, Vision Transformer (ViT), and MobileViT were initialized with pretrained weights. This significantly reduced training time and required fewer data to achieve high accuracy. Transfer learning also provided robust low-level feature extractors that were fine-tuned on the dermoscopic images, leading to improved lesion classification performance, especially in the presence of limited malignant samples.

5. Stratified Splitting

   Stratified splitting ensures that each subset of the dataset (training, validation, and testing) maintains the original class distribution. This is crucial in imbalanced datasets, like the one used in this study, where the malignant class is severely underrepresented. Without stratification, there is a risk that the validation or test set may lack sufficient malignant samples, skewing performance metrics. In this project, stratified splitting was applied during cross-validation and data partitioning, resulting in more reliable and representative evaluations of model generalization capabilities.

3.8 Model Evaluation

In clinical applications, particularly in skin cancer detection, the consequences of misclassification can be critical. A false negative—failing to identify a malignant lesion—can delay diagnosis and treatment, potentially endangering a patient's life. Therefore, evaluating model performance goes beyond general accuracy and requires metrics that reflect the model's ability to correctly identify malignant cases. This project adopted a comprehensive evaluation framework that prioritized sensitivity and minimized false negatives. The Partial Area Under the Curve (pAUC) at 80% True Positive Rate (TPR) was used as the primary metric, supplemented by a suite of additional performance indicators. To ensure the reliability and generalizability of results, k-fold cross-validation was employed as the validation strategy.

1. Partial AUC (pAUC) at 80% TPR

   The main performance metric used in this study was the Partial Area Under the Curve (pAUC) at 80% True Positive Rate (TPR). Unlike the standard ROC-AUC, which considers the model's performance across the entire spectrum of thresholds, pAUC focuses only on the region of clinical interest—where TPR exceeds 80%. This is crucial for medical diagnosis, as it emphasizes how well the model performs when high sensitivity is required. A higher pAUC in this region indicates that the model is reliable at correctly identifying malignant cases with fewer false negatives, aligning with real-world diagnostic priorities.

2. Complementary Metrics: Accuracy, Precision, Recall, F1 score, ROC-AUC

   To gain a broader understanding of the classifier's behavior, additional metrics were computed. Accuracy provides an overall sense of correctness, but can be misleading in imbalanced datasets. Precision quantifies how many predicted malignant cases are actually malignant, while recall (sensitivity) captures how many actual malignant cases were successfully detected. F1 score harmonizes precision and recall into a single

value, especially useful in scenarios with class imbalance. ROC-AUC provides an overall measure of separability between classes and is used as a standard benchmark in binary classification tasks. Together, these metrics form a holistic view of the model's diagnostic effectiveness.

3. K-Fold Cross-Validation

To ensure robustness and minimize the impact of data partitioning bias, k-fold cross-validation was applied. The dataset was split into k equally sized folds (e.g., 5 or 10), and the model was trained and evaluated k times, each time using a different fold as the validation set and the remaining folds for training. The final performance metrics were averaged across all folds, providing a more stable and reliable estimate of the model's generalization ability. This technique was particularly important given the class imbalance, helping confirm that high sensitivity and pAUC scores were not due to favorable data splits.

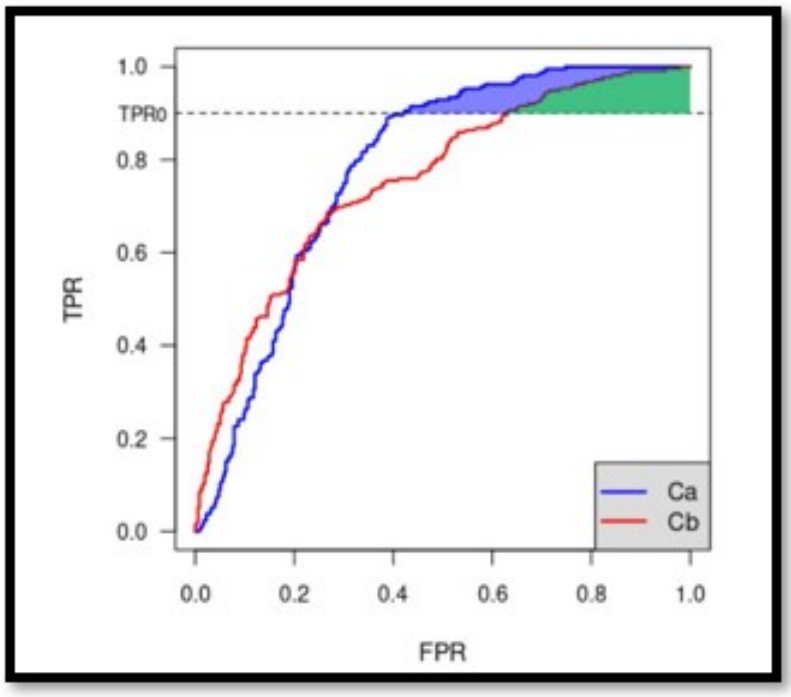

Figure 3.10: pAUC with TPR above 80%

CHAPTER 4: EXPERIMENTS

4.1 Data Pipeline

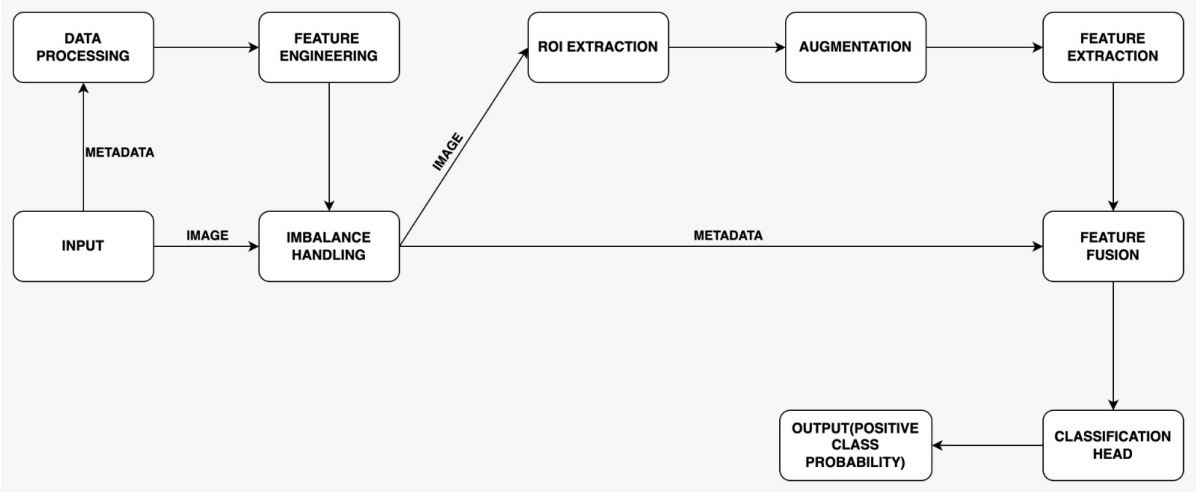

Figure 4.1: Data Pipeline

The given diagram represents a comprehensive pipeline for skin lesion classification, integrating both image-based and metadata-based processing. The workflow begins with raw input data, consisting of lesion images and associated metadata. These two data modalities undergo separate preprocessing steps before being merged later in the pipeline for classification. This approach ensures that both visual and contextual information contribute to the final prediction, improving the model's diagnostic capabilities.

Metadata undergoes an extensive data processing phase where irrelevant features, such as lesion identifiers, are removed to prevent data leakage. Categorical variables, like anatomical site and patient sex, are transformed using encoding techniques such as one-hot encoding or feature hashing, ensuring that the model can interpret them effectively. Numerical features are normalized using MinMaxScaler or StandardScaler, depending on their distribution, while dimensionality reduction techniques, such as Principal Component Analysis (PCA), help eliminate redundancy. Additionally, feature importance scores from tree-based

models guide the selection of the most predictive features, ensuring that only relevant metadata is retained for training.

In parallel, image data is processed through a region of interest (ROI) extraction step, where the lesion is isolated from surrounding skin to enhance feature clarity. To improve model robustness and generalization, data augmentation techniques are applied to the extracted ROIs. These augmentations include transformations like rotation, flipping, and contrast adjustments, mimicking real-world variations in medical imaging. Synthetic image generation methods, such as generative adversarial networks (GANs) or Mixup, are also employed to increase the diversity of malignant samples, further improving the classifier's ability to generalize.

Handling class imbalance is crucial in medical image classification, particularly when malignant cases are significantly underrepresented. Various techniques are applied to address this imbalance, including random undersampling of benign samples and synthetic oversampling of malignant ones using methods like the Synthetic Minority Over-sampling Technique (SMOTE). More advanced hybrid approaches, such as SMOTETomek and SMOTEENN, refine the synthetic sampling process by eliminating borderline or noisy samples, ensuring that the model learns from high-quality data. Additionally, class weights are incorporated into the loss function to penalize misclassification of malignant cases more heavily, emphasizing their clinical importance.

Once the data is preprocessed, features are extracted separately from the images and metadata. Convolutional Neural Networks (CNNs) such as DenseNet-169, Vision Transformer (ViT), and MedMamba are used to generate deep feature embeddings from images, capturing spatial and textural patterns indicative of malignancy. Simultaneously, tree-based models like XGBoost, LightGBM, and CatBoost process metadata, transforming patient attributes into numerical representations that complement image-based features. These extracted features encapsulate both visual and contextual information, forming a comprehensive dataset for classification.

At this stage, the extracted features from images and metadata are merged in a feature fusion step. This integration allows the classifier to leverage both modalities simultaneously, improving diagnostic accuracy. Feature fusion is a critical component of the pipeline, as it ensures that important information from both sources contributes to the final classification decision, leading to a more robust and clinically relevant prediction model.

The classification head receives the fused features and processes them through fully connected layers that refine the learned representations. LeakyReLU activation functions are used to prevent dead neurons and maintain gradient flow during training. Dropout layers are strategically added before the final output to mitigate overfitting, ensuring that the model generalizes well to unseen data. The final output is a single probability score, representing the likelihood of malignancy. This probability-based approach allows for threshold tuning, where sensitivity can be prioritized over specificity, minimizing false negatives in a clinical setting.

In summary, this pipeline integrates both image and metadata processing, incorporating imbalance-handling strategies and feature fusion techniques to enhance classification performance. The use of deep learning for image analysis and ensemble tree-based models for metadata ensures that the system captures both spatial patterns and patient-specific characteristics. By following this structured approach, the classification framework remains robust, interpretable, and well-suited for real-world medical applications.

4.2 Overview of Experimental Setup

The experimentation phase was central to our project's goal of developing an accurate, generalizable classifier for skin lesion malignancy detection. We designed a series of systematic experiments to evaluate various modeling strategies using both unimodal and multimodal data. The overarching objective was to identify which architectures and training methodologies best captured discriminative patterns from image and tabular metadata inputs. Over the course of development, we trained and validated more than 30 unique models, including a

diverse range of convolutional neural networks (CNNs), vision transformers, and ensemble tree-based methods. The models were evaluated on multiple levels, including internal train-validation splits, public leaderboard rankings, and private test performance. Metrics such as Area Under the ROC Curve (AUC), particularly on validation and private test sets, were used to benchmark generalization, while error analysis and qualitative review of predictions guided our interpretation of model behavior.

4.3 Unimodal Modeling: Image or Metadata only

4.3.1 Tree-Based Models with Metadata

We began our experimentation with traditional tree-based models—XGBoost, LightGBM, and CatBoost—applied to the tabular metadata alone. These models were advantageous for their fast training speeds and interpretability. Early results were promising in terms of training performance, with AUCs frequently reaching 1.0. However, a sharp decline in public and private leaderboard performance exposed a key limitation: severe overfitting. Even with extensive hyperparameter tuning, feature engineering, and class imbalance correction (e.g., class weighting and SMOTE), the tree-based models failed to generalize. The expressiveness of the metadata appeared insufficient by itself to support reliable malignancy classification. This highlighted the necessity of including rich visual features extracted directly from dermoscopic images.

4.3.2 CNN-Based and Transformer Models Using Dermoscopic Images

Following the limitations of metadata-only models, we shifted our focus to deep learning techniques that utilized dermoscopic images as input. CNN based architectures such as DenseNet-121, ResNet-50 were fine-tuned for the classification task. Each model was initialized with pretrained weights (e.g., ImageNet) and trained on lesion images, with classification heads consisting of dense layers, dropout, and activation functions like LeakyReLU. The improvement was immediate and substantial. For example, DenseNet-169 achieved a

validation AUC of 0.9255 and a private leaderboard score of 0.11792, significantly outperforming all metadata-only approaches. We also implemented the transformer based models such as ViT, DeIT, MobileViT and MedMamba were trained to classify based on only the images, which also produced better results than the CNN based models which is evident from the results in the table 4.1 below. These findings confirmed that image-based spatial patterns carry critical discriminative information not captured in clinical metadata.

4.4 Multimodal Fusion Models: Combining Image and Metadata

The most transformative improvement in performance came with the adoption of multimodal learning through feature fusion. In this strategy, visual features from dermoscopic images were extracted using high-capacity encoders such as DenseNet-169,ResNet 151, Vision Transformers (ViT), and MedMamba. In parallel, patient metadata underwent preprocessing and was encoded using multi-layer perceptrons (MLPs). Both the image and tabular feature vectors were then passed through dense layers for alignment before being concatenated into a single fused representation. This vector was subsequently processed through a deep neural classifier comprising several fully connected layers, dropout for regularization, and batch normalization for training stability. We also implemented Stacked ensemble models and voting classifiers which combined the predictive power of individual models such as XgBoost, LightGBM, CatBoost on the fused feature set for final classification output.

This fusion-based approach yielded superior generalization and robustness. We could see all the top performing models from the experiments turned out to be the ones with multimodal approach. This success strongly validated our hypothesis: combining spatial features with contextual metadata enables more comprehensive decision-making, especially in a clinical setting where nuanced understanding is critical. The best of the models were the stacked ensemble models that dominated all other approaches.

4.5 Hyperparameter Optimization and Training Strategies

A substantial part of our experimentation focused on refining training strategies and tuning hyperparameters. We observed that adaptive optimizers like Adam and AdamW led to faster and more stable convergence compared to traditional SGD. Particularly, AdamW offered superior generalization due to its decoupled weight decay mechanism. We also employed learning rate schedulers, such as ReduceLROnPlateau and PiecewiseConstant, to adjust learning rates dynamically based on validation loss plateaus.

Activation functions like LeakyReLU and SELU were more effective than Swish or GeLU, providing smoother gradients and preventing dead neurons. Weight initialization played a crucial role in convergence behavior, with HeNormal and Xavier Uniform leading to the most consistent training outcomes. Regularization techniques, especially dropout (ranging from 0.2 to 0.5) and batch normalization, were essential in avoiding overfitting, especially for deeper multimodal models trained on relatively limited data.

Table 4.1: Hyperparameter and Other Techniques

Category	**Best results**	**Bad results**
Optimizer Performance	Adam, AdamW (with piecewise constant learning rate schedules)	SGD (not used due to slower convergence)
Activation Functions	SELU, Leaky ReLU	Swish, GeLU
Weight Initializer	HeNormal, Xavier Initialization	Glorot uniform
Regularization & Generalization	Dropout (0.2 - 0.5), Batch Normalization	No Regularization (leads to overfitting)
Loss Functions	Binary Cross-Entropy, Focal Binary Cross Entropy Loss (with class balancing)	Cross Entropy Losses without class balancing, Squared Hinge Loss
Class Imbalance Handling	Random Undersampling & Oversampling, Class Weights	SMOTE and its variants, Near Miss
Preprocessing Techniques	KNN Imputer, One-hot encoding, Dropping columns based on semantic importance, Standard Scaler	Median, mode, ffill and bfill. Feature hashing for categorical
Callbacks Techniques	Early stopping, Saving weights with best val loss.	Reduce LR on Plateau

Loss functions were tailored for class imbalance. We found that Binary Cross-Entropy with class weights and Focal Loss significantly outperformed unweighted losses. Focal Loss, in particular, helped shift focus to harder, minority class examples and mitigated the overwhelming influence of benign cases in training.

4.6 Preprocessing and Callback Techniques

Preprocessing decisions had a notable impact on downstream model performance. For the metadata, we opted for KNN imputation to fill missing values, as it provided context-aware estimations based on feature similarity, outperforming basic mean or median strategies. Features deemed irrelevant or non-informative were dropped after analyzing feature importance via tree-based models.

Feature scaling was addressed using StandardScaler and MinMaxScaler, depending on the distribution and model requirements. For categorical features, one-hot encoding was primarily used, with feature hashing tested for high-cardinality variables. For callbacks, we implemented early stopping to terminate training once validation loss stopped improving, and model checkpointing to preserve the best weights. These techniques ensured efficient training cycles and improved reproducibility.

4.7 Insights and Key Takeaways

Our comprehensive experimentation yielded several important takeaways. Firstly, metadata alone is insufficient for accurate malignancy detection, due to its limited feature richness. Second, Image-Only Models are effective at learning distinct visual patterns for classification, but their performance eventually plateaus, irrespective of the specific architecture used (e.g., ResNet, Vision Transformer). Third, Combining image data with metadata significantly improves model performance over image-only approaches by enhancing the ability to generalize, leading to better results on unseen data and difficult-to-classify examples. Aggregating predictions from diverse models (stacking) mitigates individual model biases, im-

proves generalization, boosts performance on imbalanced datasets, reduces result variability, and increases robustness against noisy data.

We found out that among the CNN based models, Superior performance of DenseNet-169 MLP with Dense Layers, particularly in configurations with optimized training parameters. Notably, models with moderate parameter sizes (e.g., 15M–20M) achieve higher public and private test scores, suggesting an optimal trade-off between model complexity and generalization. DenseNet-169 consistently outperforms ResNet-based architectures, likely due to its efficient feature propagation and reduced redundancy in learned representations. In contrast, ResNet variants, despite their deeper architectures and higher parameter counts (e.g., ResNet-101 with 50.2M parameters and ResNetV2-50 with 75.5M parameters), exhibit lower partial AUC scores. This suggests that excessive parameterization may lead to overfitting, reducing generalization performance on unseen data.

Among the transformer based models we find that Vision Transformer MLP with 23M parameters achieves the highest public test score (0.13676), while MedMamba with 14M parameters also performs competitively with a public score of 0.13613 and a private score of 0.1228. MedMamba consistently demonstrates strong performance across different configurations, suggesting its ability to effectively capture spatial dependencies while maintaining a relatively low parameter count. Vision Transformer MLP models also perform well, particularly with parameter sizes around 10M–23M, indicating that a moderate model size leads to improved generalization. Mobile ViT MLP models, though lightweight (1M–4M parameters), exhibit slightly lower scores, highlighting the trade-off between efficiency and predictive power. Additionally, Vision Transformer MLP models with significantly reduced parameters (0.8M) show a substantial drop in performance, reinforcing the necessity of sufficient model capacity to learn meaningful representations.

Table 4.2: Tree Based Models Results

Competition Submission Name	**Models Used**	**Test - Public Score**	**Test - Private Score**	**Train AUC**	**Val AUC**
Tree based models version 9	XGBoost, LightGBM, CatBoost	0.1654	0.14565	1	0.9978
Dataset EDA – version 6	XGBoost Binary Classifier	0.1603	0.1448	1	0.9856
Dataset – EDA version 4	XGBoost	0.16416	0.13656	1	0.945
Tree based models version 12	XGBoost, LightGBM, CatBoost	0.15628	0.13445	1	0.9665
Dataset EDA version 8	XGBoost	0.15839	0.13121	0.9985	0.94
Tree based models version 10	XGBoost, LightGBM, CatBoost	0.12443	0.11055	0.9788	0.9122
dataset-eda-version 2	XGBoost	0.10538	0.07441	0.9122	0.8799
Tree based models version 11	XGBoost, LightGBM, CatBoost	0.02155	0.02104	0.9855	0.6566

Table 4.3: CNN Based Model results

Competition Submission Name	Models Used	Training Parameters	Test - Public Score	Test - Private Score	Train AUC	Val AUC
CNN based models version 32	DenseNet-169, MLP with Dense Layers	15M	0.15238	0.13614	0.9956	0.9913
Basic model building version – 21	ResNet-101, MLP with Dense Layers	50.2M	0.14285	0.12706	1	1
Basic model building version 20	ResNet_V2_50, MLP with Dense Layers	75.5M	0.1462	0.12413	0.9709	0.9198
Basic model version 24	ResNet-50, MLP with Dense Layers	439M	0.14666	0.12247	0.998	0.9485
Densenet - model	DenseNet-169, MLP with Dense Layers	14.94M	0.11977	0.11792	0.9885	0.9255
Basic model version 27	DenseNet-169, MLP with Dense Layers	20M	0.14307	0.11759	0.9944	0.997
Basic model- version 4	CNN, FCNN(Dense Layers)	26.59M	0.1164	0.10998	0.9855	0.9898
Basic model version 30	DenseNet-169, MLP with Dense Layers	20M	0.11236	0.07012	0.9269	0.9687
Basic model version 26	DenseNet-169, MLP with Dense Layers	20M	0.07654	0.05751	0.9152	0.9943
Basic model- version 15	ResNetV2-50, FCNN (Dense Layers)	25M	0.04408	0.04972	0.9999	0.9737
Basic model - version 9	ResNet-18, MLP with Dense Layers	11.7M	0.03305	0.04173	0.1681	0.6758
Basic model- version 5	ResNet-18, MLP with Dense Layers	12.8M	0.0292	0.03494	0.4566	0.5455
Basic model - version 11	ResNet-152, MLP with Dense Layers	59.5M	0.02	0.02	0.2277	0.5059
CNN based models version 31	DenseNet-169, MLP with Dense Layers	20M	0.00944	0.00947	0.9982	0.9986

Table 4.4: ViT and MedMamba Model Results

Competition Submission Name	**Models Used**	**Training Parameters**	**Test - Public Score**	**Test - Private Score**	**Train AUC**	**Val AUC**
ISIC Medmamba version 4	MedMamba	14M	0.13613	0.1228	0.9588	0.89
Vision transformers version 11	Vision Transformer, MLP	23M	0.13676	0.11722	0.9697	0.98
Vision transformers version 9	Vision Transformer, MLP	10M	0.1212	0.11688	0.9998	0.9995
ISIC Medmamba version 6	MedMamba	16M	0.13825	0.11579	0.9678	0.8445
Mobile ViT version 2	Mobile ViT, MLP	4M	0.13332	0.11469	0.9984	0.9972
ISIC Medmamba version 3	MedMamba	14M	0.12188	0.10833	0.9132	0.8234
Mobile ViT version 5	Mobile ViT, MLP	1M	0.12237	0.10822	0.9999	0.9987
Vision transformers version 2	Vision Transformer, MLP	0.8M	0.10106	0.1068	0.9963	0.9923
ISIC Medmamba version 1	MedMamba	14M	0.11892	0.10342	0.9345	0.7564
Vision transformers version 4	Vision Transformer, MLP	20M	0.08403	0.09413	0.8798	0.8928
ISIC Medmamba version 2	MedMamba	14M	0.09298	0.09374	0.8766	0.7677
Vision transformers version 5	Vision Transformer, MLP	20M	0.08228	0.09158	0.8806	0.889
Mobile ViT version 7	Mobile ViT, MLP	5M	0.04962	0.04613	1	0.9926
Vision transformers version 6	Vision Transformer, MLP	20M	0.07985	0.009617	0.9035	0.9159

CHAPTER 5: CONCLUSION

5.1 Overall Findings

This research aimed to develop a robust, accurate, and scalable deep learning model for the classification of skin lesions into malignant and benign categories using the challenging ISIC 2024 dataset. Key obstacles addressed included severe class imbalance, variability in image quality, and the implicit need for accurate lesion feature localization within images. Our investigation explored various approaches, including image-only models, multimodal architectures integrating patient metadata, and ensemble techniques.

The study revealed several key insights. Firstly, while image-only models, utilizing architectures like ResNet and Vision Transformers, effectively capture visual patterns essential for classification, their performance inherently plateaus, suggesting limitations in relying solely on visual data. Secondly, the integration of patient metadata alongside image features in a multimodal architecture proved significantly advantageous. This approach demonstrably enhanced model generalization capabilities, leading to superior performance, particularly on unseen data and samples that are inherently difficult to classify based on image appearance alone.

Furthermore, the application of ensemble methods, specifically stacked ensembles aggregating predictions from diverse models, was highly beneficial. This strategy effectively mitigated individual model biases and reduced prediction variance, leading to enhanced overall generalization. Crucially, ensembles demonstrated improved robustness against noisy data and provided a more reliable performance profile, especially pertinent given the dataset's extreme class imbalance.

Architectural comparisons yielded specific recommendations. Within CNNs, DenseNet-

169 consistently outperformed ResNet variants, particularly when optimized with moderate parameter counts (around 15M-20M), achieving a strong balance between model complexity and generalization as evidenced by top private test scores (e.g., 0.118 - 0.136). Larger ResNet models, despite higher parameter counts, tended towards lower generalization performance, suggesting a risk of overfitting. Similarly, among Transformer-based models, ViT MLP and MedMamba architectures with moderate parameter sizes (10M-23M) achieved competitive results (Private Scores 0.115 - 0.123), indicating their effectiveness. However, performance degraded with significantly smaller models (e.g., MobileViT, underparameterized ViTs), highlighting the necessity of sufficient model capacity. The multimodal deep learning approaches demonstrated the potential for high performance through combined feature learning.

This study underscores that achieving state-of-the-art performance in skin lesion classification on complex, imbalanced datasets like ISIC 2024 necessitates moving beyond image-only analysis. The integration of metadata via multimodal architectures and the strategic use of ensemble methods are crucial for enhancing generalization and robustness. Selecting appropriate base architectures (such as DenseNet or ViT/MedMamba) and carefully optimizing model complexity are key factors in developing effective diagnostic tools. These findings pave the way for more reliable AI-driven systems to aid in dermatological diagnosis.

5.2 Future Work

Based on the findings and limitations identified in this study, several promising avenues for future research emerge:

5.2.1 Advanced Multimodal Integration

- **Richer Metadata:** While patient metadata proved beneficial, future work could explore incorporating more detailed clinical information (e.g., lesion history, specific location detail, family history, dermoscopic criteria annotations if available) to potentially further

enhance model accuracy and clinical relevance.

- **Sophisticated Fusion Techniques:** Investigate more advanced fusion mechanisms beyond simple concatenation or basic attention, such as cross-modal attention transformers or graph-based methods, to better capture the complex interplay between visual features and metadata.

5.2.2 Addressing Data Challenges

- **Advanced Imbalance Handling:** Although ensembles helped mitigate imbalance, explore dedicated techniques like cost-sensitive learning, specialized loss functions (e.g., Focal Loss, Balanced Softmax), or generative adversarial networks (GANs) for synthetic minority oversampling specifically tailored to skin lesion data.

- **Image Quality Enhancement & Domain Adaptation:** Develop preprocessing pipelines or incorporate domain adaptation techniques specifically designed to handle the variability in image quality (lighting, resolution, occlusions) within the ISIC dataset and potentially improve robustness to images from different sources or clinical settings.

- **Longitudinal Data Analysis:** If available, incorporating patient follow-up data or images of the same lesion over time could provide valuable temporal information for improved diagnostic accuracy and prognostic prediction.

5.2.3 Model Architecture and Ensemble Refinement

- **Hybrid Architectures:** Explore hybrid models that combine the strengths of CNNs (local feature extraction) and Transformers/State Space Models like Mamba (global context modeling) within a single architecture for potentially superior feature representation.

- **Optimized Ensemble Strategies:** Investigate alternative or more efficient ensemble methods, such as snapshot ensembles or Bayesian model averaging, and explore techniques to optimize ensemble diversity and reduce computational overhead.

- **Explicit Feature Localization:** While classification models implicitly learn localization, integrating explicit segmentation modules or attention mechanisms (like Grad-CAM) more deeply into the training process could improve model focus on relevant lesion areas and enhance interpretability.

5.2.4 Interpretability and Explainability (XAI)

- **Clinical Validation of Explanations:** Apply state-of-the-art XAI techniques (e.g., SHAP, LIME, attention maps) to the best-performing multimodal and ensemble models. Critically, validate whether the features highlighted by these methods align with clinical diagnostic criteria used by dermatologists, thereby increasing trust and potential for clinical adoption.

5.2.5 Clinical Translation and Deployment

- **Prospective Clinical Validation:** Evaluate the performance of the developed models in a prospective clinical setting using real-world, unseen data collected under routine conditions to assess true diagnostic utility.

- **Efficiency and Deployment:** Optimize the best-performing models (potentially using techniques like knowledge distillation or quantization) for deployment in resource-constrained environments, such as mobile devices or web applications, to facilitate point-of-care use.

- **Multi-Class Classification:** Extend the models beyond binary classification to differentiate between various subtypes of benign and malignant lesions (e.g., melanoma, basal cell carcinoma, nevus, seborrheic keratosis), providing more granular diagnostic information.

These future directions aim to build upon the successes of integrating multimodal data and ensemble techniques, further address the inherent challenges of the ISIC dataset, and move towards developing clinically reliable and interpretable AI tools for dermatological diagnosis.

REFERENCES

[1] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li, “Deep learning-based image segmentation on multimodal medical imaging,” *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, no. 2, pp. 162–169, Mar. 2019.

[2] T. Zhou, S. Ruan, and S. Canu, “A review: Deep learning for medical image segmentation using multi-modality fusion,” *Array*, vol. 3–4, p. 100004, 2019.

[3] Taleb, A., Lippert, C., Klein, T. & Nabi, M. Multimodal Self-supervised Learning for Medical Image Analysis. *Information Processing In Medical Imaging*. pp. 661-673 (2021)

[4] M. A. Kassem, K. M. Hosny, and M. M. Fouad, “Skin lesions classification into eight classes for ISIC 2019 using deep convolutional neural network and transfer learning,” *IEEE Access*, vol. 8, pp. 114822–114832, 2020.

[5] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, “Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC),” in *Proc. IEEE 15th Int. Symp. Biomed. Imaging (ISBI)*, 2018, pp. 168–172.

[6] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li, ”Deep learning-based image segmentation on multimodal medical imaging,” *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, no. 2, pp. 162–169, 2019.

[7] P. Luo, Y. Li, L.-P. Tian, and F.-X. Wu, ”Enhancing the prediction of disease–gene associations with multimodal deep learning”, *Bioinformatics*, vol. 35, no. 19, pp. 3735–3742, Oct. 2019. doi:

https://doi.org/10.1093/bioinformatics/btz15510.1093/bioinformatics/btz155.

[8] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *Proc. 2018 IEEE 15th Int. Symp. Biomed. Imaging (ISBI)*, pp. 168–172, 2018. doi: https://doi.org/10.1109/ISBI.2018.836354710.1109/ISBI.2018.8363547.

[9] M. A. Kassem, K. M. Hosny, and M. M. Fouad, "Skin lesions classification into eight classes for ISIC 2019 using deep convolutional neural network and transfer learning," *IEEE Access*, vol. 8, pp. 114822–114832, 2020. doi: https://doi.org/10.1109/ACCESS.2020.300389010.1109/ACCESS.2020.3003890.

[10] A. Taleb, C. Lippert, T. Klein, and M. Nabi, "Multimodal self-supervised learning for medical image analysis," in *Information Processing in Medical Imaging (IPMI 2021)*, A. Feragen, S. Sommer, J. Schnabel, and M. Nielsen, Eds., Lecture Notes in Computer Science, vol. 12729, Springer, Cham, 2021. doi: https://doi.org/10.1007/978-3-030-78191-0$_5$110.1007/$978-3-030-78191-0\_51$.

[11] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017.

[12] N. Codella et al., "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging," *IEEE TMI*, vol. 38, no. 10, pp. 2520–2532, 2019.

[13] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset," *Scientific Data*, vol. 5, 180161, 2018.

[14] L. Bi et al., "Dermoscopic image segmentation via multistage fully convolutional networks," *IEEE TMI*, vol. 36, no. 10, pp. 2069–2081, 2019.

[15] L. C. Chen et al., "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, pp. 801–818, 2018.

[16] Q. Abbas, M. E. Celebi, and I. F. Garcia, "Hair removal methods: A comparative study for dermoscopy images," *Biomed. Signal Process. Control*, vol. 6, no. 4, pp. 395–404, 2011.

[17] I. Gonzalez-Diaz, "DermaKNet: Incorporating the knowledge of dermatologists to convolutional neural networks," *Comput. Methods Programs Biomed.*, vol. 187, 105242, 2019.

[18] A. Vaswani et al., "Attention is all you need," in *NeurIPS*, pp. 5998–6008, 2017.

[19] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[20] H. Touvron et al., "Training data-efficient image transformers distillation through attention," in *ICML*, 2021.

[21] S. Mehta et al., "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint arXiv:2110.02178*, 2021.

[22] T. J. Brinker et al., "A convolutional neural network trained with dermatoscopic images performed on par with 145 dermatologists," *Eur. J. Cancer*, vol. 111, pp. 148–154, 2019.

[23] T. C. Pham et al., "Deep CNN and data augmentation for skin lesion classification," in *IEEE EMBC*, pp. 2610–2613, 2018.

[24] M. A. Al-Masni et al., "Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks," *Comput. Methods Programs Biomed.*, vol. 162, pp. 221–231, 2018.

[25] Y. Liu et al., "A survey of deep learning-based object detection," *Pattern Recognit.*, vol. 121, 108152, 2020.

[26] M. Raghu et al., "Transfusion: Understanding transfer learning for medical imaging," in *NeurIPS*, pp. 3347–3357, 2019.

[27] Q. Abbas et al., "Pattern classification of dermoscopy images: A perceptually uniform model," *Pattern Recognit.*, vol. 47, no. 1, pp. 110–120, 2014.

[28] B. Harangi, "Skin lesion classification with ensembles of deep convolutional neural networks," *J. Biomed. Inform.*, vol. 86, pp. 25–32, 2018.

[29] A. Mahbod et al., "Fusing fine-tuned deep features for skin lesion classification," *Comput. Med. Imaging Graph.*, vol. 71, pp. 19–29, 2021.

[30] F. Xie et al., "Melanoma classification using dynamic feature fusion from deep convolutional neural networks," *IEEE Access*, vol. 8, pp. 150524–150534, 2020.

[31] A. Bissoto et al., "Deep learning and data augmentation for skin lesion classification," in *ISIC Challenge*, 2019.

[32] S. Chaturvedi et al., "Deep learning-based skin cancer classification using hybrid feature fusion," *Comput. Biol. Med.*, vol. 140, 105097, 2022.

[33] T. Saba et al., "Segmentation of skin lesions and classification using deep learning," *Neural Comput. Appl.*, vol. 33, pp. 13183–13193, 2021.

[34] N. Gessert et al., "Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss," *IEEE TMI*, vol. 39, no. 5, pp. 1485–1494, 2020.

[35] Y. Li et al., "Dense anatomical attention network for skin lesion classification," *IEEE Access*, vol. 7, pp. 88359–88371, 2019.

[36] J. Zhang et al., "Meta-learning with global-class prototypes for few-shot skin disease classification," *Med. Image Anal.*, vol. 73, 102173, 2022.

[37] Y. Yuan et al., "Automatic skin lesion segmentation with fully convolutional-deconvolutional networks," *arXiv preprint arXiv:1703.05165*, 2017.

[38] F. Xie et al., "Multiclass skin lesion classification using a hybrid deep learning approach," *IEEE JBHI*, vol. 20, no. 4, pp. 1026–1036, 2016.

[39] M. A. Khan et al., "Skin lesion segmentation and classification: A survey," *IEEE Access*, vol. 10, pp. 16590–16621, 2022.

[40] S. Pathan et al., "Skin lesion classification using convolutional neural network," *Procedia Comput. Sci.*, vol. 132, pp. 434–441, 2018.

[41] J. Zhang et al., "Attention residual learning for skin lesion classification," *IEEE JBHI*, vol. 23, no. 2, pp. 547–558, 2019.

[42] L. Wang et al., "A hybrid model for melanoma detection using deep learning," *IEEE Access*, vol. 7, pp. 90359–90370, 2019.

[43] S. Reshma and M. Krishnan, "Classification of skin lesions using deep CNN with extensive data augmentation," *Procedia Comput. Sci.*, vol. 171, pp. 1174–1181, 2020.

[44] P. Tang et al., "Self-supervised pretraining of transformer networks for skin cancer classification," in *MICCAI*, pp. 499–509, 2021.

[45] W. Zhu et al., "Skin lesion diagnosis using multi-modal fusion of dermoscopic and clinical images," *IEEE Access*, vol. 9, pp. 9603–9615, 2021.

[46] A. Mahbod et al., "Transfer learning using a deep convolutional neural network for skin lesion classification," *Comput. Methods Programs Biomed.*, vol. 197, 105600, 2020.

[47] S. Albahli et al., "Deep learning for skin cancer detection and classification: A comprehensive review," *Multimed. Tools Appl.*, vol. 81, pp. 11171–11201, 2022.

[48] M. Goyal et al., "Multi-class skin disease classification using deep residual networks," in *Medical Imaging 2017: Image Processing*, vol. 10133, 2017.

APPENDIX A: Appendix

Find the below image of the custom deep convolutional network architecture that combines metadata for the classification head.

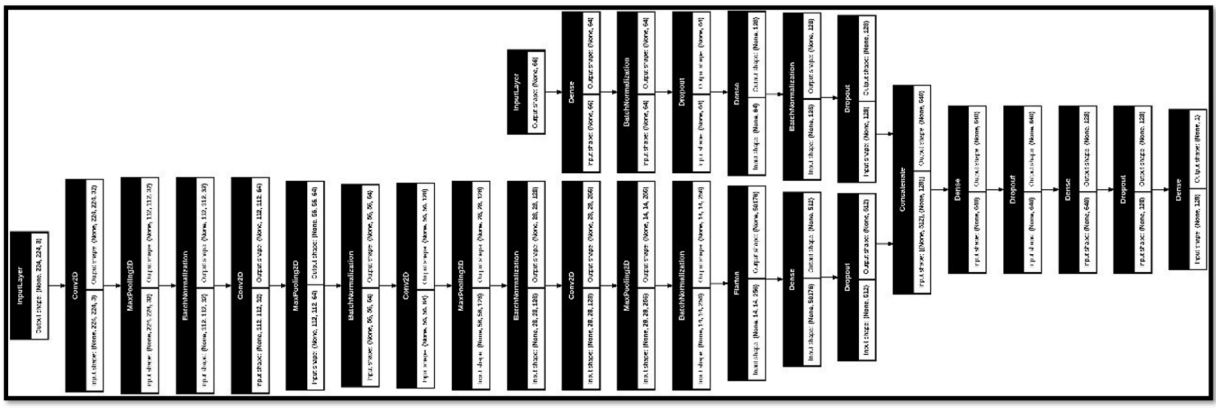

Figure A.1: CNN Architecture