People's Democratic Republic of Algeria
الجمهورية الجزائرية الشعبية الديموقراطية
Ministry of Higher Education and scientific Research
وزارة التعليم العالي والبحث العلمي

University of Kasdi Merbah Ouargla
Faculty of New Information Technologies and Communication
Department of Computer Science and Information Technology

Thesis Submitted in Candidacy for a Master's Degree in Computer Science

Specialty: Artificial Intelligence and Data Science

Presented by:

Khaldi Mouad & Khaldi Abdelhamid

Supervised by : **Prof. Merieme Khelifa & Prof. Mezzoudj Saliha**

Early Detection of Skin Cancer Using A Deep Learning Method and Optimization Methods

Publicly Defended on **June 12h 2025** at **08h00**

Board of Examiners:

Prof. Bensaci Ramla	Kasdi Merbah University	President
Prof. Khaldi Belal	Kasdi Merbah University	Examiner

Academic Year: 2024/2025

Acknowledgments

Firstly and foremost, all praise, gratitude, and acknowledgement are due to Allah, the Most Gracious, the Most Merciful. We are profoundly thankful to Him, the Almighty, for granting us the strength to persevere, the wisdom to understand, and the guidance to navigate the challenges of this research. It is through His infinite blessings and mercy that we were provided with the means, opportunities, and the ability to undertake this endeavor and bring it to fruition. We acknowledge that all success comes from Him alone, and we are humbled by His continuous support and the causes He provided throughout our academic journey, leading to this achievement. He is indeed the most deserving of all thanks.

Following this profound gratitude, we extend our deepest appreciation to our supervisor, Professor Mezzoudj Saliha. Her meticulous supervision, insightful guidance, and unwavering commitment to ensuring this work was conducted as properly as possible were invaluable. We are truly grateful for her mentorship and dedication throughout this process.

We would also like to thank the members of the thesis committee for their time and forthcoming feedback during the examination process.

Our sincere acknowledgements go to several professors whose teaching and support during our master's studies in Artificial Intelligence and Data Science were particularly impactful. We are especially grateful to Professor Aiadi Oussama, Professor Khaldi Belal, Professor Kherfi Mohammed Lamine, and Professor Khadra Bouanane for their invaluable and indispensable guidance. Their technical expertise and the personal impact they had on shaping our views within the field were profoundly appreciated during these past two years.

We are also indebted to all the professors and instructors who have taught and guided us throughout our entire university academic journey, from our first year up until this final master's year. Each contributed significantly to the foundation upon which this work is built.

On a personal level, we owe a profound debt of gratitude to our family. To our wonderful parents and our dear sister, thank you for your tremendous and unconditional support, encouragement, and patience. You have been along for the ride since day one, long before university began, and your belief in us has been a constant source of strength. Special thanks are truly deserved.

Finally, we wish to extend special thanks to our dear friend, Babasidi Mohammed Khaled. We are grateful for his friendship, for being part of this journey, providing unwavering support, and cheering for our success all along.

Completing this thesis would not have been possible without the initial and ultimate blessing of Allah, followed by the collective support, guidance, and encouragement of all these individuals. Thank you.

Dedication

This thesis is lovingly dedicated to our cherished family,
whose unwavering belief has been our guiding light:

To our wonderful parents and our precious sister,
Thank you for your constant love and steadfast support,
a presence felt every single day, guiding us
from our very first steps into school
all the way to this final culmination of our academic journey.
Your love has been our anchor.

To our dear grandmother and our uncles,
Your enduring encouragement and wishes for us
to reach the highest ceilings we could possibly aspire to
have always motivated us to strive further.
Thank you for believing in our potential.

Your collective love, support, and aspirations are woven into every page of this work.

Table of Contents

List of Tables

List of Figures

Abbreviations

AI : Artificial Intelligence

AUC : Area Under the Curve

ANN : Artificial Neural Network

CAD : Computer-Aided Diagnosis

CNN : Convolutional Neural Network

CPU : Central Processing Unit

CT : Computed Tomography

DL : Deep Learning

DNN : Deep Neural Network

FN : False Negative

FP : False Positive

GAN : Generative Adversarial Network

GMM Gaussian Mixture Model

GPU : Graphics Processing Unit

IoU : Intersection over Union

ISIC : International Skin Imaging Collaboration

LBP : Local Binary Pattern

LSTM : Long Short-Term Memory

ML : Machine Learning

MSER : Maximally Stable Extremal Region

NLP : Natural Language Processing

OvR : One-vs-Rest

PCA : Principal Component Analysis

ReLU : Rectified Linear Unit

RF : Random Forest

RL : Reinforcement Learning

ROC : Receiver Operating Characteristic

RNN : Recurrent Neural Network

SAM : Segment Anything Model

SK : Seborrhoeic Keratosis

SVM : Support Vector Machine

TN : True Negative

TP : True Positive

UV : Ultraviolet

ViT : Vision Transformer

Abstract: The early and accurate diagnosis of skin cancer, particularly malignant melanoma, is critical for patient survival, yet it poses significant challenges due to the visual complexity of dermoscopic images and inherent dataset imbalances. This thesis investigates and develops a comprehensive pipeline for automated skin lesion classification using the ISIC 2017 dataset. The methodology encompasses a rigorous multi-stage preprocessing workflow, including hair removal and image enhancement, followed by an extensive data augmentation strategy to create a balanced training set (4116 images) from the originally imbalanced data. The study comparatively evaluates multiple segmentation techniques (U-Net, SAM) and classification strategies. The primary classification approach involves using a pre-trained EfficientNetB7 model as a deep feature extractor from whole images, with subsequent classification performed by various machine learning models, including Support Vector Machines (SVM) and Random Forests. Experimental results on a validation split of the balanced data showed promising performance, achieving up to 84.1% accuracy. However, a critical performance drop to approximately 66% accuracy was observed on the independent, imbalanced ISIC 2017 test set, with recall for the crucial melanoma class falling below 10%. This study concludes that while robust preprocessing and data balancing can yield high performance in a controlled setting, significant challenges in model generalization persist. The performance gap highlights that domain shift and a lack of robustness to real-world class imbalance are major obstacles, underscoring the need for advanced techniques to create clinically reliable automated diagnostic systems.

Keywords: AI, Machine Learning, Skin Cancer Classification, Deep Learning, Dermoscopic Image Analysis, ISIC 2017, EfficientNetB7.

Résumé: Le diagnostic précoce et précis du cancer de la peau, en particulier du mélanome malin, est essentiel pour la survie des patients, mais il présente des défis importants en raison de la complexité visuelle des images dermoscopiques et des déséquilibres inhérents aux ensembles de données. Cette thèse étudie et développe un pipeline complet pour la classification automatisée des lésions cutanées en utilisant l'ensemble de données ISIC 2017. La méthodologie comprend un flux de travail de prétraitement rigoureux en plusieurs étapes, incluant la suppression des poils et l'amélioration de l'image, suivi d'une stratégie extensive d'augmentation de données

pour créer un ensemble d'entraînement équilibré (4116 images) à partir des données initialement déséquilibrées. L'étude évalue de manière comparative plusieurs techniques de segmentation (U-Net, SAM) et stratégies de classification. L'approche de classification principale consiste à utiliser un modèle EfficientNetB7 pré-entraîné comme extracteur de caractéristiques profondes à partir d'images entières, la classification ultérieure étant effectuée par divers modèles d'apprentissage automatique, y compris les Machines à Vecteurs de Support (SVM) et les Forêts Aléatoires. Les résultats expérimentaux sur un sous-ensemble de validation des données équilibrées ont montré des performances prometteuses, atteignant jusqu'à 84,1 % de précision. Cependant, une chute de performance critique à environ 66 % de précision a été observée sur l'ensemble de test indépendant et déséquilibré d'ISIC 2017, avec un rappel pour la classe cruciale du mélanome tombant en dessous de 10 %. Cette étude conclut que, bien qu'un prétraitement robuste et un équilibrage des données puissent donner des performances élevées dans un cadre contrôlé, des défis importants en matière de généralisation du modèle persistent. L'écart de performance met en évidence que le décalage de domaine (domain shift) et le manque de robustesse au déséquilibre de classe du monde réel sont des obstacles majeurs, soulignant la nécessité de techniques avancées pour créer des systèmes de diagnostic automatisés cliniquement fiables.

Mots-clés : IA,Apprentissage automatique,Classification du Cancer de la Peau, Apprentissage Profond, Analyse d'Images Dermoscopiques,ISIC 2017 EfficientNetB7.

ملخص: يعد التشخيص المبكر والدقيق لسرطان الجلد، وخاصة الميلانوما الخبيثة، أمراً حاسماً لبقاء المرضى (الدرموسكوبي)على قيد الحياة، ولكنه يطرح تحديات كبيرة بسبب التعقيد البصري للصور الجلدية بالتنظير والاختلالات الكامنة في مجموعات البيانات. تبحث هذه الأطروحة وتطور خط عمل شامل للتصنيف الآلي لآفات الجلد باستخدام مجموعة بيانات ISIC 2017. تشمل المنهجية المتبعة عملية معالجة أولية دقيقة متعددة المراحل، بما في ذلك إزالة الشعر وتحسين الصور، تليها استراتيجية واسعة لزيادة البيانات لإنشاء مجموعة تدريب متوازنة (صورة 4116) من البيانات الأصلية غير المتوازنة. تقيّم الدراسة بشكل مقارن تقنيات متعددة للتجزئة (U-Net, SAM) واستراتيجيات للتصنيف. يتمثل نهج التصنيف الأساسي في استخدام نموذج EfficientNetB7 مُدرَّب مسبقًا لاستخلاص الميزات العميقة من الصور الكاملة، مع إجراء التصنيف لاحقًا بواسطة نماذج تعلم الآلة المختلفة، بما في ذلك آلات المتجهات الداعمة (SVM) والغابات العشوائية. أظهرت النتائج التجريبية على عينة التحقق من صحة البيانات المتوازنة أداءً واعدًا، حيث بلغت الدقة 84.1%. ومع

ISIC 2017عند التقييم على مجموعة اختبار 66%ذلك، لوحظ انخفاض حاد في الأداء إلى ما يقرب من تخلص هذه .10%المستقلة وغير المتوازنة، مع انخفاض نسبة استدعاء فئة الميلانوما الحاسمة إلى أقل من الدراسة إلى أنه على الرغم من أن المعالجة الأولية القوية وموازنة البيانات يمكن أن تحققا أداءً عاليًا في بيئة يسلط هذا الفارق في الأداء الضوء على .خاضعة للرقابة، إلا أن تحديات كبيرة في تعميم النموذج لا تزال قائمة ونقص المتانة في مواجهة عدم توازن الفئات في العالم الحقيقي يمثلان (domain shift)أن تغير النطاق عقبات رئيسية، مما يؤكد الحاجة إلى تقنيات متقدمة لإنشاء أنظمة تشخيص آلية موثوقة سريريًا.

الكلمات المفتاحية: تصنيف سرطان الجلد ISIC2017,التعلم الالي ،تحليل الصور الجلدية التعلم العميق، بالتنظير ، EfficientNetB7.

General Introduction

Medical imaging stands as a cornerstone of modern medicine, offering invaluable, non-invasive insights for disease detection, diagnosis, and treatment monitoring. The sheer volume of medical image data generated daily underscores the critical need for sophisticated computational tools to aid in its processing and interpretation. Within this landscape, skin cancer represents a significant global health concern, ranking among the most common cancers worldwide. For instance, in 2022, non-melanoma skin cancer was reported as the 5th most diagnosed malignancy globally with over 1.2 million new cases, while melanoma, the most dangerous form, was the 17th, contributing over 330,000 new cases **[1]** . These conditions arise from the uncontrolled growth of abnormal skin cells, often linked to cumulative exposure to ultraviolet (UV) radiation but possible anywhere on the body. Accurately identifying malignant lesions through visual inspection of dermoscopic images remains a complex task due to the wide variation in lesion appearance and the subtle differences that can exist between benign and cancerous growths.

The imperative for early detection of skin cancer cannot be overstated, as it directly correlates with patient survival rates and treatment success. Melanoma, in particular, highlights this urgency: the 5-year survival rate when melanoma is localized to the skin (Stage 0 or Stage I) is reported to be very high, often approaching 98-100% . However, if diagnosis is delayed and the cancer metastasizes (Stage IV), this rate can plummet dramatically to figures around 22.5% **[2]** , although advancements in treatment continue to offer hope. Globally, skin cancer also imposes a significant mortality burden; in 2022, non-melanoma types were estimated to cause approximately 69,000 deaths and melanoma around 59,000 deaths **[1]**. The inherent difficulty in visual differentiation, even for trained dermatologists, combined with inherent challenges in medical datasets – such as the significant class imbalance observed in benchmarks like the ISIC 2017 dataset used in this study (where benign cases vastly outnumber malignant ones) – strongly motivates the development and refinement of automated systems to support clinical decision-making and improve diagnostic accuracy.

The quest for improved diagnostic accuracy has led to significant interest in computer-aided diagnosis (CAD) systems, particularly those leveraging Artificial Intelligence (AI). Over the past decade, Deep Learning (DL), a subfield of AI, has revolutionized medical image analysis. Convolutional Neural Networks (CNNs), with architectures drawing inspiration from the human visual cortex, have demonstrated a remarkable ability to automatically learn intricate patterns and discriminative features directly from pixel data, often reducing the need for laborious manual feature engineering . Numerous studies, frequently utilizing large, annotated datasets such as those from the International Skin Imaging Collaboration (ISIC) , have showcased the potential of DL in skin lesion classification. Various advanced CNN architectures (like VGG, ResNet, and EfficientNet, and newer models like Vision Transformers (ViT) have achieved impressive results in specific research contexts. Indeed, some studies published in the literature have reported classification accuracies exceeding 90%, and in certain specialized tasks or using particular datasets like ISIC, figures as high as 95.6% on ISIC 2017, or even up to 97-99.5% in specific configurations, have been documented **[3]**. While these results showcase the immense potential of DL, they also highlight a critical gap. Translating high performance from controlled experimental setups to robust generalization on unseen, real-world clinical data, especially with inherent dataset imbalances and domain shifts, remains a significant research challenge. This sets the stage for the central research problem, or *problematic*, that this thesis confronts: Given the known challenges of class imbalance and domain shift, how can we develop and evaluate a deep learning pipeline for skin lesion classification that is not only accurate in a balanced training environment but also robust and generalizable to real-world, imbalanced clinical data ? This thesis explores this question in detail by systematically building, testing, and analyzing a comprehensive pipeline, from data preprocessing and augmentation to feature extraction and classification, ultimately revealing the profound impact of these challenges on model performance.

This Master's thesis addresses the critical challenge of early skin cancer detection and classification through the application and rigorous evaluation of computational intelligence techniques. Our research investigates a comprehensive pipeline leveraging the internationally recognized ISIC 2017 dataset. The core contributions of this thesis are centered on:

1. The systematic implementation and evaluation of **extensive data preprocessing and augmentation strategies** specifically designed to address the significant class imbalance inherent in the ISIC 2017 training data.
2. A comparative analysis of various **skin lesion segmentation methods**, including traditional MSER, the deep learning-based U-Net, a hybrid MSER+U-Net approach, and the foundational Segment Anything Model (SAM), to understand their efficacy in delineating lesion boundaries.
3. An in-depth investigation into different **feature extraction and classification paradigms**. This includes exploring features from SAM-segmented regions versus whole-image features, comparing traditional LBP texture features with deep features from EfficientNetB7, and evaluating the performance of multiple machine learning classifiers (SVM, Random Forest, Logistic Regression) with hyperparameter optimization.
4. Critically, this work provides a detailed analysis of the **challenges in generalizing models** from a balanced, augmented training environment to an imbalanced, unseen test set, offering insights into the impacts of domain shift and class distribution disparities.

The overarching goal remains the development of a more robust understanding of the factors influencing automated skin lesion classification accuracy, thereby contributing to the ongoing efforts to build reliable systems for dermatological assessment.

This manuscript is organized into the following chapters:

- **Chapter 1 :** provides the essential medical context for this research. It introduces the structure of the skin, defines the relevant types of skin cancer, and describes the clinical diagnostic pathway, with a focus on dermoscopy as the primary imaging modality.
- **Chapter 2 :** This chapter provides a comprehensive review of existing research on skin cancer detection using deep learning, with a primary focus on the ISIC 2017 dataset. It examines seminal contributions, benchmark challenges, and state-of-the-art models that tackle critical issues such as class imbalance, lack of generalization to unseen data, and architectural limitations in traditional convolutional neural networks. Particular attention is given to

advanced methods including patch-based perceptual attention models, multi-scale feature fusion, ensemble learning strategies, and hybrid optimization techniques. Through detailed analysis and comparison of major works, this chapter identifies key gaps in the literature and sets the foundation for the methodological choices and innovations proposed in this thesis.

- **Chapter 3 :** outlines the specific methodology employed in this research. This includes a detailed description of the ISIC 2017 dataset, the data preprocessing pipeline (including hair removal, denoising, enhancement, and resizing), the data augmentation strategy for class balancing, the architecture and training of the deep learning models used for segmentation (U-Net) and classification (EfficientNetB7-based approaches), the implementation of traditional methods (MSER, LBP), the use of foundation models (SAM), and the complete experimental design for both segmentation and classification tasks.
- **Chapter 4 :** reports and discusses the experimental results obtained. It includes a thorough evaluation of model performance for both segmentation (MSER, U-Net, MSER+U-Net, SAM) and classification (initial hybrid model, EfficientNetB7 feature extraction with various ML classifiers, attempted attention-enhanced EfficientNetB7) using standard metrics. This chapter also provides a comparative analysis, discusses the significant performance discrepancies between validation and test sets, analyzes the reasons for these differences (domain shift, class imbalance, augmentation effects), and addresses computational challenges and limitations of the study.
- **General Conclusion** encapsulates the main findings and contributions of the thesis, reiterates the answers to the research objectives, and proposes potential avenues for future research stemming from this work.

Chapter 1: Introduction To Skin Cancer & its Detection

1.1 Introduction:

Before delving into the computational methods for automated skin cancer detection, it is essential to first establish a firm understanding of the clinical problem domain. This chapter provides the necessary medical and dermatological context that motivates this entire thesis. We begin by exploring the basic biology of the skin and defining the different types of skin cancer relevant to this study, particularly melanoma, seborrheic keratosis, and nevus. Subsequently, we will describe the standard clinical diagnostic pathway and focus on dermoscopy, the non-invasive imaging technique that has become the cornerstone for the visual assessment of skin lesions and the source of the data used in this research. This foundational knowledge is crucial for appreciating the complexities and challenges involved in visually differentiating benign from malignant lesions, thereby setting the stage for the computational approaches reviewed and developed in the subsequent chapters.

1.2 Understanding Skin Cancer:

1.2.1 The Skin: A Brief Overview:

The skin is the human body's largest organ, serving crucial functions including protection against external threats (pathogens, UV radiation), temperature regulation, and sensory perception **[4]** . It is structurally composed of three primary layers:

Epidermis: The outermost layer, providing the main protective barrier. It contains various cell types, including keratinocytes (the most abundant), melanocytes (producing melanin pigment), Langerhans cells (immune function), and Merkel cells (touch sensation).

Dermis: The middle layer, located beneath the epidermis, provides structural support, elasticity, and houses blood vessels, nerves, hair follicles, and glands. It's primarily composed of connective tissue (collagen, elastin).

Hypodermis (Subcutaneous Tissue): The innermost layer, consisting mainly of adipose (fat) tissue, providing insulation, energy storage, and cushioning.

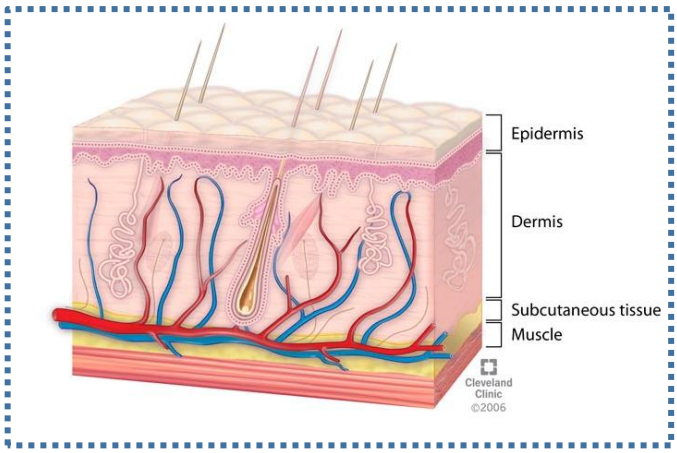


Figure 1 : Diagram Ilustrating Layers of Skin

1.2.2 Skin Cancer: Definition and Significance:

Skin cancer arises from the uncontrolled proliferation of abnormal cells within one of the skin's layers, primarily the epidermis. It is the most common type of cancer globally, with incidence rates continuing to rise in many parts of the world . While often associated with excessive exposure to ultraviolet (UV) radiation from the sun or artificial tanning beds, other factors like genetics, skin type, and immune status also play a role. The significance of skin cancer lies not only in its high prevalence but also in the potential morbidity and mortality associated with its aggressive forms, particularly melanoma .

1.2.3 Relevant Types of Skin Lesions:

Skin lesions can be broadly categorized as benign (non-cancerous) or malignant (cancerous). This thesis, utilizing the ISIC 2017 dataset, focuses primarily on distinguishing between the following common types:

Melanoma: This is the most dangerous form of skin cancer, arising from melanocytes. While less common than other types like basal cell carcinoma (BCC) or squamous cell carcinoma (SCC), melanoma is responsible for the vast majority of skin cancer deaths due to its high potential to metastasize (spread) to other parts of the body if not detected early . Dermoscopically, melanomas often exhibit asymmetry, irregular borders, multiple colors (brown, black, red, blue, white) **[5]**, and changing structures over time. Specific features can include atypical pigment networks, streaks, pseudopods, blue-white veil, and irregular vascular patterns .

Figure 2 : Example of Melanoma Skin Lesion

Seborrheic Keratosis (SK): These are very common, benign (non-cancerous) growths that often appear as waxy, slightly raised lesions, looking as if "stuck on" the skin surface. They can vary in color from light brown to black . While harmless, their appearance, especially when pigmented, can sometimes mimic melanoma, making accurate differentiation important. Dermoscopy typically reveals characteristic features like milia-like cysts (small white or yellowish spots), comedo-like openings (dark plugs), fissures, ridges, and a 'cerebriform' (brain-like) pattern **[6]**. They often have a sharply demarcated border .

Figure 3 : Example of Seborrhoeic Keratosis Skin Lesion

Nevus (Mole): Nevi are common benign growths of melanocytes. They can be present at birth (congenital) or acquired later in life. Most nevi are harmless, but some atypical nevi can resemble early melanoma, and rarely, melanoma can arise within an existing nevus. Dermoscopy of benign nevi often shows symmetric structures, regular borders, and uniform patterns **[7]**, such as a reticular (network) pattern, globular pattern, or homogeneous (structureless) pattern, depending on the

nevus type and depth . Specific patterns exist for different body sites (e.g., parallel furrow pattern on palms/soles) .

1.2.4 Risk Factors and the Imperative of Early Detection:

While UV radiation is the primary environmental risk factor, others include fair skin type, history of sunburns, multiple atypical moles, family history of skin cancer, and weakened immune system. The critical importance of early detection, especially for melanoma, cannot be overemphasized. As highlighted in the General Introduction, survival rates are significantly higher when melanoma is diagnosed and treated in its early stages (approaching 100 % 5-year survival for Stage 0/I) compared to late-stage diagnoses where the cancer has spread (dropping to around 22.5% for Stage IV) . This stark difference underscores the need for effective screening and accurate diagnostic tools.

1.3 Clinical Diagnosis and Imaging:

1.3.1 Standard Diagnostic Pathway:

The process usually begins with a clinical skin examination by a primary care physician or dermatologist. Suspicious lesions are identified based on visual characteristics (e.g., the "ABCDE" rule - Asymmetry, Border irregularity, Color variation, Diameter > 6mm, Evolving/Changing). If a lesion is deemed suspicious, the gold standard for definitive diagnosis is a skin biopsy, where part or all of the lesion is removed and examined histopathologically under a microscope.

1.3.2 Dermoscopy:

Dermoscopy (also known as dermatoscopy or epiluminescence microscopy) is a non-invasive imaging technique that has become a cornerstone of clinical dermatology for evaluating skin lesions. It utilizes a handheld magnification device (dermatoscope) with a specific light source (often polarized or using immersion fluid) to illuminate the skin and eliminate surface reflection, allowing visualization of subsurface structures within the epidermis and superficial dermis that are not visible to the naked eye **[8]** . This enhanced visualization significantly improves diagnostic accuracy for both pigmented and non-pigmented lesions compared to standard clinical examination alone. Trained clinicians use dermoscopy to identify

specific morphological features, patterns, and colors associated with benign or malignant conditions (as described for melanoma, SK, and nevus in Section (1.2.3) . Diagnostic algorithms like the 7-point checklist or pattern analysis are often employed in conjunction with dermoscopy. The images used in major research datasets like ISIC are typically high-resolution dermoscopic images .

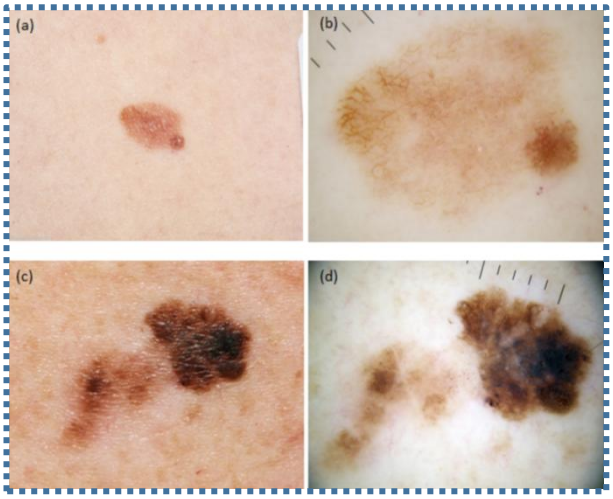


Figure 4 : Two Examples of Macroscopic Images (a and c) and Dermoscopic Images (b and d) : (a) and (b) Are Images of a Dysplastic Nevus and, (c) and (d) Are of an Invasive Melanoma

1.3.3 Other Imaging Modalities:

While dermoscopy is the most widely used imaging aid in clinical practice for pigmented lesions, other advanced non-invasive imaging techniques exist and are subjects of research, including:

- **Reflectance Confocal Microscopy (RCM):** Provides real-time, high-resolution images of skin layers at a cellular level, akin to "optical biopsy".
- **Optical Coherence Tomography (OCT):** Creates cross-sectional images of skin layers, similar to ultrasound but using light.
- **High-Frequency Ultrasound:** Can assess lesion depth and structure.
- **Spectroscopy Techniques:** Analyze the interaction of light with tissue to infer biochemical composition.

While powerful, these techniques are often more specialized, expensive, or less widely available than dermoscopy, which remains the focus for large-scale dataset collection and CAD development relevant to this thesis.

1.4 Computer-Aided Diagnosis (CAD) for Skin Lesion Analysis:

The visual complexity of dermoscopic images and the subtle differences between lesion types present challenges even for experienced dermatologists. This has motivated the development of Computer-Aided Diagnosis (CAD) systems to provide objective, quantitative analysis and assist clinicians in decision-making, potentially improving diagnostic accuracy, consistency, and efficiency.

1.4.1 Traditional Machine Learning Approaches:

Early attempts at CAD for skin lesion classification relied on traditional machine learning pipelines. These typically involved several distinct steps:

1. **Preprocessing:** Image enhancement, noise reduction (e.g., hair removal) and sometimes lesion segmentation (manually or semi-automatically isolating the lesion from surrounding skin).
2. **Feature Extraction:** Manually defining and calculating quantitative features based on clinical knowledge or image properties. Common features aimed to capture aspects of the ABCD(E) criteria, such as color variance, texture measures (e.g., using Gray Level Co-occurrence Matrix - GLCM), shape descriptors (e.g., border irregularity index, asymmetry measures), and structural features.
3. **Classification:** Training standard machine learning classifiers (e.g., Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Decision Trees, Random Forests) using the extracted features to predict the lesion type (e.g., benign vs. malignant).

While these traditional methods, which rely on hand-engineered features, offer the distinct advantage of greater **interpretability**—allowing clinicians to understand which specific features (e.g., border irregularity, color asymmetry) are driving the classification—their overall efficacy is often constrained by several critical dependencies. First, their performance is fundamentally limited by the **quality and**

relevance of the pre-selected features, a process that requires extensive domain expertise and may fail to capture the full, complex range of patterns present in the lesion. Second, these systems are highly susceptible to **error propagation**; inaccuracies in an initial, mandatory segmentation step can irreversibly corrupt the feature extraction process and lead to incorrect classifications. Finally, this approach lacks the ability to **automatically discover novel, more discriminative features** directly from the image data, a key advantage offered by end-to-end deep learning models that learn feature hierarchies automatically.

1.4.2 Deep Learning Approaches:

The advent of Deep Learning (DL), particularly Convolutional Neural Networks (CNNs), marked a paradigm shift in image analysis, including medical imaging. CNNs have the remarkable ability to learn hierarchical representations of features directly from raw pixel data through layers of convolution, non-linearity, and pooling operations, eliminating the need for manual feature engineering . This end-to-end learning capability has led to state-of-the-art performance in various computer vision tasks.

a) Application to Skin Lesion Classification:

Researchers rapidly applied CNNs to skin lesion classification using dermoscopy images, often leveraging large, publicly available datasets like those curated by the International Skin Imaging Collaboration (ISIC). ISIC has hosted annual challenges since 2016, providing benchmark datasets (including ISIC 2017, the focus of this work) and fostering rapid progress in the field.

b) Key Architectures and Techniques:

Numerous studies have explored various DL architectures and techniques:

CNN Architectures: Standard, well-established architectures like VGG **[9]** (e.g., VGG-16, VGG-19) , ResNet **[10]** (e.g., ResNet-50) , Inception (e.g., Inception-v3) , DenseNet , and MobileNet have been widely adapted, often using transfer learning. This involves taking a model pre-trained on a large general image dataset (like ImageNet) and fine-tuning it on the smaller, specific medical dataset (like ISIC). This approach leverages the general feature extraction capabilities learned

from millions of images and often leads to better performance than training from scratch, especially with limited medical data . More recent architectures like Vision Transformers (ViT) are also being explored.

Data Augmentation: To combat limited dataset size and increase model robustness, extensive data augmentation is commonly applied. This involves creating modified versions of training images through operations like rotation, flipping, scaling, shifting, brightness/contrast adjustments, and elastic deformations.

Handling Class Imbalance: ISIC datasets are often highly imbalanced (many more benign nevi than melanomas or SKs). A model trained on such data will be biased towards the majority class, leading to poor detection of rare but critical minority classes. Various techniques are employed to specifically counteract this problem:

- **Data-level approaches:** This is the most direct method to address imbalance. Techniques like **oversampling** (e.g., duplicating minority class images or using synthetic data generation like SMOTE) or **undersampling** (removing majority class images) directly alter the training set distribution. This forces the model to see a more balanced representation of each class during training, preventing it from simply learning to ignore the minority classes.
- **Algorithm-level approaches:** These methods modify the learning algorithm itself to give more importance to the minority classes. Using **weighted loss functions** is a common example. This addresses imbalance by applying a higher penalty for misclassifying an example from a minority class, thereby forcing the model to pay much closer attention to getting those predictions right.
- **Ensemble Models:** While generally used to improve performance, ensembles can be specifically designed to tackle class imbalance. One such technique is training multiple different models on various **balanced subsets** of the data (e.g., each subset contains all samples of the minority class but only a small, different fraction of the majority class). By combining the predictions from these "experts," the final decision is less biased by the overall majority class than a single model would be, effectively improving minority class recognition.
- **Segmentation-then-Classification:** This is an indirect approach. By first using a model like U-Net to precisely segment the lesion, it removes distracting

background information from the image. This helps the subsequent classifier focus only on the relevant lesion features. This can indirectly help with class imbalance by making the subtle, discriminative features of minority class lesions more prominent and easier for the classifier to learn, rather than being lost in noise.

- **Attention Mechanisms:** These modules can also indirectly mitigate the effects of imbalance. An attention mechanism guides the model to focus on the most salient regions of an image. For a rare lesion class whose distinguishing features are subtle and localized, attention helps the model "zoom in" on these critical areas. This prevents the model from being distracted by more common, but less important, features shared across all classes, thus improving its ability to identify the unique patterns of the minority class.
- **Optimization and Loss Functions:** This is a powerful algorithm-level approach. Beyond standard optimizers, designing **custom loss functions** is a direct way to address imbalance. For instance, a loss function like the **MDKLoss** was specifically designed to enhance the separation between lesion types by incorporating medical domain knowledge, forcing a greater distance in the feature space between classes that are visually similar but clinically different, which is often the case for rare malignant lesions versus common benign ones.

c) Performance on ISIC Datasets (including ISIC 2017)

Studies utilizing ISIC datasets consistently demonstrate the high potential of DL. While direct comparison is complex due to variations in preprocessing, augmentation, training protocols, specific dataset splits, and evaluation metrics, reported performance figures are often impressive:

* Accuracies frequently exceed 90-85% for multi-class classification tasks

* Area Under the Receiver Operating Characteristic Curve (AUC), a common metric for binary classification (e.g., melanoma vs. non-melanoma), often surpasses 0.90, with some studies reporting AUCs above 0.95 .

* Specific studies using ISIC 2017 have reported high results, although the exact figures vary. For example, one paper references a table showing classification results on the ISIC 2017 test set , while another mentions achieving an AUC of 81.4% on ISIC 2017 using a specific clustering approach . Others report accuracies like 95.6% on ISIC 2017 data within broader experiments . Ensemble methods on balanced ISIC

data have reported classification accuracies reaching 99%

It is crucial, however, to interpret these results cautiously, considering the specific task (binary vs. multi-class), the exact dataset partition used (official challenge test set vs. custom splits), and the potential for overfitting or dataset bias.

1.4.3 Common Challenges in CAD for Skin Cancer:

Despite significant progress, several challenges remain in developing robust and clinically reliable CAD systems for skin cancer detection:

Data Limitations: While ISIC datasets are invaluable, they can still be considered limited compared to datasets used in general computer vision. Acquiring large, diverse, expertly annotated medical datasets is difficult due to privacy concerns, cost, and time required for expert labeling .

Class Imbalance: The natural prevalence of lesion types leads to imbalanced datasets, making it harder for models to learn effective representations for rare but critical classes like melanoma .

Image Quality and Variability: Dermoscopic images can vary significantly in quality due to different devices, lighting conditions, operator skill, and the presence of artifacts like hair, air bubbles, ink markings, or low contrast . Models need to be robust to these variations.

Intra-class Variability & Inter-class Similarity: Lesions within the same class can appear very different (high intra-class variability), while lesions from different classes can look very similar (low inter-class similarity), making discrimination challenging .

Generalization: Models trained on specific datasets may not generalize well to images from different populations, equipment, or clinical settings. Ensuring robustness and avoiding dataset bias is critical.

Interpretability: Many DL models function as "black boxes," making it difficult to understand why they make a particular prediction. For clinical acceptance, improving model interpretability and providing clinicians with insights into the decision process is important.

1.5 Chapter Conclusion:

This chapter has provided the essential clinical and medical foundation for the research presented in this thesis. We have reviewed the basic structure of the skin, defined the key characteristics of melanoma, seborrheic keratosis, and benign nevi, and outlined the standard diagnostic process. The critical role of dermoscopy in enhancing the visualization of subsurface lesion structures has been highlighted, establishing it as the primary imaging modality that enables advanced computational analysis. It is clear that despite the utility of dermoscopy, significant visual overlap between benign and malignant conditions creates a persistent diagnostic challenge. This challenge of accurate and consistent visual interpretation forms the primary motivation for developing computer-aided diagnosis systems. The next chapter will critically review the scientific literature on existing computational methods designed to tackle this very problem.

Chapter 2: Literature Review

2.1 Introduction

A thorough literature review is foundational in any scientific endeavor, especially in rapidly evolving fields such as medical image analysis and deep learning. In the context of skin cancer detection, the literature review serves to map out existing methodologies, highlight their strengths and limitations, and identify areas of improvement that this research aims to address. Reviewing previous work provides not only a contextual background but also a methodological framework that shapes the formulation of the problem, choice of dataset, model architecture, and evaluation strategy.

This chapter explores the major contributions and limitations in skin cancer detection using dermoscopic image analysis, focusing primarily on studies based on the International Skin Imaging Collaboration (ISIC) datasets, especially ISIC 2017. It systematically analyzes studies that address issues of class imbalance, generalization to unseen data, and architectural innovations in convolutional neural networks (CNNs), attention mechanisms, ensemble learning, and hybrid optimization techniques.

2.2 The ISIC 2017 Challenge: A Benchmark Foundation

The **ISIC 2017 challenge**, as documented by Codella et al. (2018), was a landmark effort to create a common benchmark for the dermoscopic analysis of skin lesions**[11]**. The challenge focused on three main tasks:

- **Task 1: Lesion Segmentation** — Participants developed models to delineate lesion boundaries in dermoscopic images. The top-performing method achieved a Jaccard Index of 0.765, Dice coefficient of 0.849, and pixel-wise accuracy of 93.4%, using a variation of a fully convolutional network (FCN) ensemble.

-

Task 2: Dermoscopic Feature Detection — This task required detection of four clinical features: pigment network, negative network, streaks, and milia-like cysts. The SLIC algorithm was used to segment images into superpixels for classification. The top model achieved an AUC of 0.895 (average), with individual category scores ranging from 0.807 to 0.960.

-

Task 3: Disease Classification — Participants classified images into three classes: melanoma (374 training samples), seborrheic keratosis (254), and benign nevi (1372). The top average performer achieved an overall AUC of 0.911, with individual class AUCs of 0.868 (melanoma) and 0.953 (seborrheic keratosis).

Key Issues Identified:

-

Severe Class Imbalance: The ISIC 2017 dataset suffers from significant class imbalance, with benign nevi making up over 70% of the samples. This skew heavily biases models towards majority class predictions.

-

Generalization Gap: Many top-performing models were shown to fail on external datasets due to overfitting and lack of diversity in the training set.

-

Lack of Ethnic and Device Diversity: Images were predominantly from light-skinned individuals and captured using specific dermoscopy devices, limiting generalizability.

This challenge was the first to standardize and publish large-scale dermoscopic data and evaluation protocols, significantly advancing reproducibility and comparability in the field.

2.3 Addressing Data Imbalance and Generalization in ISIC 2017

Menegola et al. (2017) — Ensemble and Class-Aware Learning

Menegola et al., participants in ISIC 2017, explored various ensemble techniques and data augmentation methods to mitigate the class imbalance **[12]**. They constructed an ensemble of deep CNNs trained with cost-sensitive loss functions and performed targeted augmentations for underrepresented classes.

Despite these strategies, they observed that the performance on melanoma remained low compared to benign classes. Their analysis indicated that the model's confusion matrix showed a high rate of false negatives for melanoma, emphasizing that even ensemble learning with augmentation could not fully mitigate imbalance without a more fundamental dataset restructuring.

Codella et al. (IBM Research, 2018) — Probabilistic Score Normalization and Fusion

Codella et al. proposed probabilistic score normalization to standardize outputs from multiple models, enabling more meaningful ensemble learning. They introduced three fusion strategies: average score fusion, linear SVM (L-SVM), and nonlinear SVM (NL-SVM). Among these, L-SVM achieved the highest overall classification performance with an AUC of 0.926 and melanoma-specific AUC of 0.892.

Their analysis showed that fusion of classifiers improved sensitivity and specificity balance for melanoma, particularly at critical sensitivity thresholds (82%, 89%, and 95%). However, their models still struggled with generalization to novel image domains not represented in the ISIC 2017 data.

Fishbaugh et al. — Rank Aggregation for Robust Performance

Fishbaugh et al. implemented a rank aggregation method to combine outputs from heterogeneous models. Their strategy emphasized consensus-building between classifiers rather than simple averaging. They found that models with complementary error patterns improved ensemble performance significantly, particularly for melanoma and SK.

Their results provided early evidence that hybrid model consensus can partially address imbalance by leveraging different models' strengths on minority classes.

2.4 Patch-Based Learning with Perceptual Attention

Nayak et al. (2024) — Patch-Based CNN with Spatial Attention

Nayak et al. addressed the limitations of resizing dermoscopic images by proposing a patch-based ensemble approach that retained high-resolution feature detail **[13]**. Their architecture, based on MobileNetV2, extracted features from multiple patches (whole image, 2-patch, 4-patch settings) and fused them into a composite intermediate feature matrix.

To maintain the spatial integrity of patch-based features, they introduced a perceptual attention mechanism inspired by NLP attention models. This module aligned features across patches, preventing semantic drift and enabling coherent classification.

Key Contributions:

- Demonstrated that increasing the number of patches improved classification accuracy (whole: 82.57%, 2-patch: 85.36%, 4-patch: 88.28%).
- Achieved superior results over standard resizing methods that lost critical edge and texture features.
- Trained on a balanced subset of ISIC 2017 (374 melanoma and 374 nevus), avoiding dataset bias.

Their approach serves as a baseline for attention-integrated, patch-aware CNNs in dermoscopic classification.

2.5 Multi-Scale Feature Fusion with Hybrid Optimization

Prakash et al. (2025) — MFFDCNN-CTDC Architecture

Prakash et al. developed the **Multi-Scale Feature Fusion of Deep CNNs for Cancerous Tumor Detection and Classification (MFFDCNN-CTDC) [14]**, a robust pipeline combining segmentation, feature fusion, classification, and parameter optimization.

Architecture Details:

- **Preprocessing**: Applied Sobel filters for noise reduction and edge enhancement.
- **Segmentation**: Used UNet3+, a densely connected version of UNet with lateral and residual connections for high-resolution semantic segmentation.
- **Feature Extraction**: Combined ResNet50 and EfficientNet to capture features across spatial scales and network depths.
- **Classification**: Employed a Convolutional Autoencoder (CAE) to model nonlinear feature interactions.
- **Optimization**: Fine-tuned CAE parameters using a hybrid Fireworks Whale Optimization Algorithm (FWWOA).

Results:

-

Achieved 98.78% accuracy on ISIC 2017 and 99.02% on HAM10000.

-

Outperformed other state-of-the-art methods in AUC, sensitivity, and specificity.

-

Demonstrated robustness against noisy inputs and rare lesion types.

Their work represents one of the most comprehensive deep learning systems for skin cancer classification and is particularly relevant for handling multi-class imbalances and image variability.

2.6 Summary of Models and Comparative Analysis

Table 1: Models Comparative Table

Study	Dataset	Method	Key Metrics	Strengths	Weaknesses
Codella et al. (ISIC 2017)	ISIC 2017	Multi-task CNN + fusion	AUC = 0.926 (L-SVM)	Ensemble fusion, standardized benchmark	Dataset imbalance, generalization issues
Menegola et al.	ISIC 2017	Ensemble + Augmentation	AUC < 0.89 (melanoma)	Class-aware training, diverse models	Poor melanoma precision
Nayak et al. (2024)	ISIC 2017 (subset)	Patch-based MobileNetV2 + Attention	Acc = 88.28%	High spatial retention, low computation	Binary classification only
Prakash et al. (2025)	ISIC + HAM10000	UNet3+ + ResNet + CAE + FWWOA	Acc = 98.78%	Multi-scale fusion, robust	Training complexity

				optimization	
Fishbaugh et al.	ISIC 2017	Rank aggregation ensemble	Not specified	Hybrid ensemble, balanced voting	Implementation complexity

2.7 Critical Gaps and Future Directions

Despite notable progress, key challenges remain:

- **Class Imbalance Solutions**: There is a need for more advanced strategies such as SMOTE for image generation, class-weighted loss functions, and focal loss.

- **Dataset Diversity**: Current datasets underrepresent darker skin tones and rare lesion types. Future datasets must be demographically and clinically diverse.

- **Model Generalization**: Few studies validate models on cross-dataset setups. Domain adaptation techniques like CycleGANs and adversarial training remain underutilized.

- **Clinical Interpretability**: The majority of models are black-box systems. Incorporating explainability (e.g., Grad-CAM, SHAP) can enhance clinical trust.

2.8 Conclusion

This chapter presented a comprehensive review of recent advancements in skin cancer detection using deep learning, with a focus on ISIC 2017. It detailed the evolution from benchmark datasets to patch-based learning, hybrid optimization, and ensemble strategies. Through analysis of major works, it became clear that while accuracy has improved significantly, the field continues to grapple with issues of imbalance, bias, and generalization.

These insights directly inform the motivation and methodology of this research, which aims to develop a lightweight yet accurate, attention-enhanced model that generalizes well across diverse patient populations and clinical settings.

Chapter 3: Materials and Methodology

3.1 Introduction

This chapter details the comprehensive methodology employed in this research to investigate and develop automated systems for skin lesion segmentation and classification. The primary goal, as outlined in Chapter 1, is the effective analysis of dermoscopic images from the ISIC 2017 dataset, leveraging a combination of traditional image processing techniques, established deep learning architectures, and emerging foundation models. Adhering to scientific rigor, this chapter will meticulously describe the materials used, the multi-phase experimental design, and the specific procedures undertaken at each stage, from initial data acquisition and extensive preprocessing to the implementation and evaluation of various segmentation and classification models.

The methodology is presented sequentially, reflecting the iterative nature of the research process. We begin by describing the dataset and the computational environment. This is followed by a detailed account of Phase 1: data preprocessing and augmentation, essential for preparing the data for model training. Phase 2 details the exploration of different lesion segmentation techniques, including MSER, U-Net, and SAM. Phase 3 describes the approaches taken for feature extraction and the classification of lesions into benign, melanoma, and seborrheic keratosis categories, including initial hybrid models and subsequent refined strategies focusing on EfficientNetB7, culminating in the current work on an attention-enhanced fine-tuning approach. Each step is described with sufficient detail to allow for reproducibility, a core tenet of scientific research. The choice of specific methods and parameters is justified based on the requirements of the task and insights gained throughout the experimental process.

3.2 Materials:

3.2.1 Dataset: ISIC 2017 The primary dataset utilized in this study is the **ISIC (International Skin Imaging Collaboration) 2017 Challenge dataset**. This publicly available dataset is a well-recognized benchmark for developing and evaluating algorithms for skin lesion analysis, particularly dermoscopic images. For this research, all operations were performed on a local copy of the dataset, named "Copy Of ISIC2017," to ensure the preservation of the original data integrity.

The ISIC 2017 dataset comprises several tasks; for this thesis, we primarily focused on:

- **Training Data:**
  - ISIC-2017_Training_Data: Contains original dermoscopic images of skin lesions.
  - ISIC-2017_Training_Part1_GroundTruth: Contains corresponding binary segmentation masks for the training images.
  - ISIC-2017_Training_Part3_GroundTruth.csv: Provides the classification labels for the training images. An initial analysis revealed a significant class imbalance:
    - Benign (Nevus): 1372 samples
    - Melanoma: 374 samples
    - Seborrheic Keratosis (SK): 254 samples This imbalance was a key challenge addressed in the preprocessing phase.

- **Test Data:**

  - ISIC-2017_Test_v2_Data: Contains 600 test images for classification.
  - ISIC-2017_Test_v2_Part1_GroundTruth: Contains segmentation ground truth for the test images (used for evaluating segmentation models).
  - ISIC-2017_Test_v2_Part3_GroundTruth.csv: Contains the classification ground truth for the test images (used for final classification model evaluation).

The dataset contains RGB color dermoscopic images of varying sizes and captured under different conditions, presenting realistic challenges for automated analysis.

3.2.2 Development Environment and Tools:

The implementation of the methodologies described in this thesis was carried out using the Python programming language, leveraging several key open-source libraries and frameworks widely used in the machine learning and computer vision communities.

- **Programming Language:** Python (version 3.12.6).
- **Core Libraries:**
    - **OpenCV (cv2):** Extensively used for image processing tasks such as image reading/writing, color space conversions, morphological operations, filtering (Gaussian, Bilateral), inpainting, and image resizing.
    - **Keras (with TensorFlow backend):** The primary deep learning framework used for building, training, and evaluating neural network models, including the U-Net and EfficientNetB7 architectures. ImageDataGenerator from Keras was specifically used for data augmentation.
    - **Scikit-learn:** Employed for traditional machine learning classifiers (SVM, Logistic Regression, Random Forest), performance evaluation metrics (accuracy, precision, recall, F1-score, confusion matrix), and hyperparameter tuning (GridSearchCV).
    - **NumPy & Pandas:** For numerical operations, array manipulations, and handling CSV metadata files.
- **Specialized Models/Tools:**
    - **Segment Anything Model (SAM):** The pre-trained ViT-B backbone of SAM was utilized for promptable segmentation experiments.
- **Computational Resources:** Training deep learning models, especially phases involving fine-tuning large architectures like EfficientNetB7, can be computationally intensive. While not explicitly detailed in the provided work summary, it is implied that adequate computational resources (potentially including GPUs) were necessary for these tasks, particularly highlighted by the long training times reported for Phase 2 of the current fine-tuning work.

3.3 Phase 1: Data Preprocessing and Augmentation:

This initial phase was foundational, aiming to clean the raw ISIC 2017 training images, standardize their properties, and prepare them for effective model training. The need for preprocessing arises from common artifacts and variations in

dermoscopic images that can hinder algorithmic performance. All operations were performed on the copied dataset to preserve the original data.

3.3.1 Sequential Image Processing Pipeline: A multi-step pipeline was applied to each image in the ISIC-2017_Training_Data set:

- **Step 1: Hair Removal (DullRazor Variant)**
  - **Rationale:** Hair is a common artifact in dermoscopic images that can obscure lesion details or be misinterpreted as part of the lesion structure by algorithms.
  - **Method:**

1. The input BGR image was converted to Grayscale to simplify hair detection.
2. A Morphological Black-hat filter (using a 9x9 rectangular kernel) was applied. The black-hat operation highlights dark objects on a lighter background, making it suitable for detecting thin, dark hair strands.
3. The output of the black-hat filter was smoothed using a 3x3 Gaussian blur to reduce minor noise in the detected hair regions.
4. The blurred result was then thresholded (threshold value set to 10) to create a binary mask isolating the detected hair structures.
5. OpenCV's inpaint function (using the TELEA algorithm with an inpainting radius of 6 pixels) was employed to fill in the regions defined by the hair mask, using information from surrounding pixels to create a visually plausible reconstruction of the underlying skin.

- **Output:** Images with removed hair were saved to the Hair_Removed_Images folder.

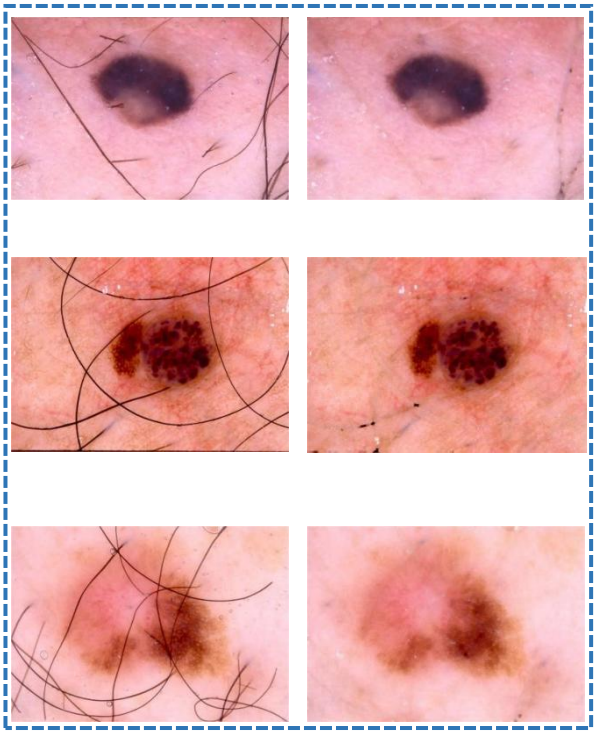


Figure 5 : Example demonstrating images before and after hair removal

- **Step 2: Noise Removal (Bilateral Filtering):**
  - **Rationale:** Dermoscopic images can contain various types of noise. It's important to reduce this noise while preserving the sharpness of important lesion edges, which are crucial for diagnosis.
  - **Method:**

    Bilateral filtering (cv2.bilateralFilter) was applied to the hair-removed images. The parameters used were: diameter d=9, sigmaColor=75, and sigmaSpace=75. The bilateral filter is an edge-preserving smoothing filter; it considers both the spatial proximity of pixels and their intensity/color similarity, thus smoothing flat regions while retaining sharp edges.

- **Output:** Denoised images were saved to the Denoised_Images folder.

- **Step 3: Image Enhancement (CLAHE + Gamma Correction):**
  - **Rationale:** To improve the visibility of lesion features by enhancing local contrast and adjusting overall image brightness, making the images more suitable for subsequent analysis.
  - **Method:**

1. The input BGR image was converted to Grayscale to simplify hair detection.

2. A Morphological Black-hat filter (using a 9x9 rectangular kernel) was applied. The black-hat operation highlights dark objects on a lighter background, making it suitable for detecting thin, dark hair strands.

  - **Output:** Enhanced images were saved to the Enhanced_Images folder.

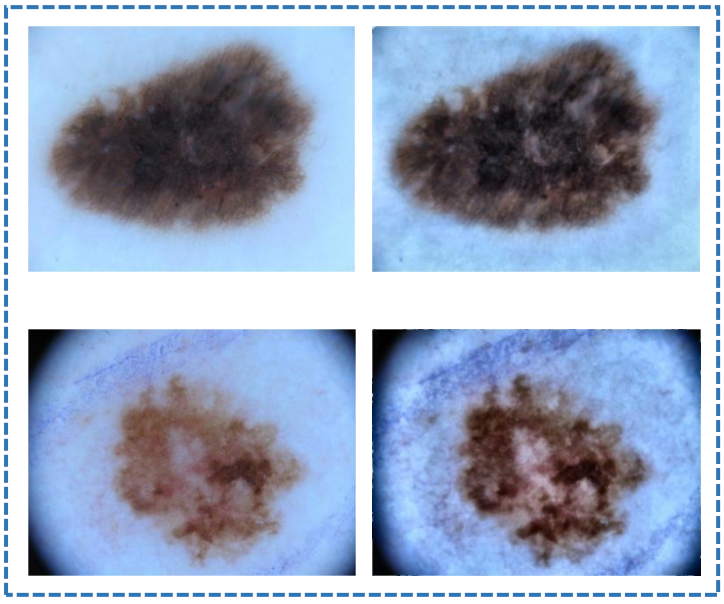


Figure 6 :Example demonstrating images after enhancement and contrast adjustment

- **Step 4: Image Resizing**
    - **Rationale:** Deep learning models typically require input images of a fixed, standardized dimension.
    - **Method:**
      All images output from the enhancement step were resized to 224x224 pixels using cv2.resize. Linear interpolation was likely used for images to maintain quality. Corresponding ground truth segmentation masks were also resized to 224x224 pixels, typically using nearest-neighbor interpolation to preserve the discrete mask labels (0 or 255).
    - **Output:** Resized images were saved to Resized_Images, and resized masks were stored separately in a Resized Training Ground Truth folder.

3.3.2 Data Augmentation for Class Balancing

- **Rationale:** The ISIC 2017 training dataset exhibits a significant class imbalance (Benign: 1372, Melanoma: 374, SK: 254). Training models on such imbalanced data can lead to bias towards the majority class (Benign), resulting in poor performance on minority classes (Melanoma, SK), which are often of critical diagnostic importance.
- **Strategy:** To mitigate this, an oversampling strategy was adopted for the minority classes (Melanoma and SK) through data augmentation. The goal was to create a balanced training set where each class had 1372 samples.
- **Implementation Details:**
    1. **Data Organization:** The 224x224 preprocessed images (from Resized_Images) and their corresponding resized masks were initially sorted into class-specific subfolders: Melanoma, Seborrheic_Keratosis, and Benign.
    2. **Augmentation Generators (Keras ImageDataGenerator):**
        - For **images**: Random transformations were applied, including rotation (±180°), width and height shifts (±10% of total dimension), zoom (range of ±10%), horizontal and vertical flips, and brightness adjustments (range of ±10%). The

fill_mode='nearest' was used to fill pixels outside the input boundaries after transformations.

- For **masks**: Identical geometric transformations (rotation, shifts, zoom, flips) were applied to ensure correspondence with the augmented images. However, brightness adjustments were excluded. fill_mode='constant' with cval=0 (black) was used to preserve the integrity of the binary mask regions.

3. **Synchronized Augmentation:** A custom function (augment_pair) was implemented to ensure that each image and its corresponding mask underwent the exact same geometric transformations. This was achieved by using a shared random seed for both the image and mask augmentation generators for each pair. After augmentation, masks were re-binarized (thresholded at a value like 127, assuming original mask values are 0 and 255) to ensure they remained binary (0 or 255).
4. **Oversampling Process:** The augment_class function iteratively applied the augmentation to the Melanoma and Seborrheic Keratosis classes until each class contained 1372 samples (original + augmented). Augmented images were saved to Augmented_Images/[Class_Name] and augmented masks to Augmented_Masks/[Class_Name].
5. **Final Balanced Dataset Assembly:** The original 1372 preprocessed Benign images/masks were combined with the original and newly augmented Melanoma and SK images/masks. This resulted in the final balanced training dataset stored in Final_Training_Images and Final_Training_Masks, each containing 4116 samples (1372 samples × 3 classes).
6. **Metadata Update:** The original ISIC-2017_Training_Part3_GroundTruth.csv file was updated to include entries for all the augmented images, resulting in a new Updated_Training_Part3_GroundTruth.csv file reflecting the balanced 4116 samples.

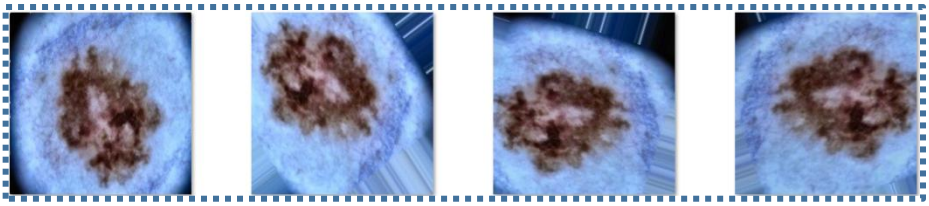


Figure 7 : Example of an original lesion image and augmented images corresponding to it

3.4 Phase 2: Segmentation Methodologies:

This phase explored various computational techniques to accurately segment the lesion area from the surrounding healthy skin in dermoscopic images. Effective segmentation is often a critical precursor to accurate feature extraction and classification.

3.4.1 Maximally Stable Extremal Regions (MSER)

- **Concept:** MSER is a feature detection algorithm that identifies stable connected components (regions) in an image by analyzing how their properties change across a range of intensity thresholds. These regions are "extremal" because all pixels inside have either higher (bright MSERs) or lower (dark MSERs) intensity than all pixels on their outer boundary. "Maximally stable" refers to regions whose area changes minimally over a range of thresholds.
- **Implementation (Default Parameters):** Initially, MSER was applied to grayscale test images using default OpenCV parameters: delta=5, min_area=60, max_area=14400, max_variation=0.25, min_diversity=0.2. Binary segmentation masks were generated by computing and filling the convex hulls of the detected MSER regions.
- **Implementation (Optimized Parameters):** Recognizing the limitations of default parameters, a separate Genetic Algorithm (GA) optimization process had been used to find MSER parameters that maximized IoU on a training subset. These optimized parameters were: delta=6, min_area=638, max_area=21889, max_variation=0.27875, min_diversity=0.47989. These

optimized parameters were subsequently used for standalone evaluation and as a component in a hybrid model.

3.4.2 U-Net Architecture for Segmentation:

- **Concept:** The U-Net architecture is a deep convolutional neural network specifically designed for biomedical image segmentation **[15]**. Its symmetrical encoder-decoder ("U" shape) structure with skip connections (concatenating feature maps from the encoder path to the decoder path) allows it to capture both context and precise localization, making it highly effective.
- **Implementation:** A standard U-Net model was constructed using the Keras library.
    - **Input:** 224x224x3 RGB images from the Final_Training_Images (balanced dataset).
    - **Output:** A 224x224x1 probability map, where each pixel value represents the probability of that pixel belonging to the lesion class. A Sigmoid activation function was used in the final layer for this purpose.
- **Training:** The U-Net was trained on the 4116 balanced preprocessed images and their corresponding masks (Final_Training_Images and Final_Training_Masks).
    - **Optimizer:** Adam optimizer.
    - **Loss Function:** Binary Cross-Entropy, suitable for binary segmentation tasks.
    - The model achieving the best performance on a validation split (implicitly, though not explicitly detailed how the split was made for U-Net training in the document) was saved as unet_checkpoint_light.h5.

3.4.3 MSER+U-Net Hybrid Model

- **Concept:** This approach aimed to leverage the strengths of both MSER and U-Net by guiding the U-Net with an initial segmentation proposal from MSER.
- **Implementation:**
    1. Optimized MSER (using the GA-derived parameters) was applied to generate binary masks.
    2. These MSER masks were resized to 224x224 and normalized (likely to range).
    3. The normalized MSER mask was concatenated with the original RGB image channels, creating a 4-channel input (R, G, B, MSER-mask) for a modified U-Net architecture (adjusted to accept 4 input channels).
- **Training:** The hybrid model was trained similarly to the standard U-Net on the balanced dataset. The best performing model was saved as best_model.h5.

3.4.4 Segment Anything Model (SAM) - ViT-B Backbone

- **Concept:** SAM is a state-of-the-art foundation model for image segmentation, designed to be promptable and capable of zero-shot generalization to unseen objects and image distributions **[16]**.
- **Implementation:** The pre-trained ViT-B (Vision Transformer - Base) backbone of SAM was utilized.
    - For initial evaluation and its use in the first hybrid classification algorithm, segmentation was performed on test images by loading the RGB image and then manually providing a bounding box prompt around the lesion area using OpenCV's cv2.selectROI function.
    - The SamPredictor object then generated a binary mask based on this user-provided prompt.
    - This method demonstrates SAM's zero-shot capabilities when guided by a simple prompt.

3.5 Phase 3: Classification Methodologies

Following the segmentation explorations, this phase focused on classifying lesions into one of three categories: Benign (Nevus), Melanoma, or Seborrheic Keratosis (SK), primarily using feature extraction from images.

3.5.1 Initial Hybrid Classification Approach (Inspired by Teacher's Algorithm 1)

- **Concept:** This approach aimed to combine information from a foundation model for segmentation (SAM), traditional texture features (LBP), and deep features from a powerful CNN (EfficientNetB7), followed by a traditional ML classifier (SVM).
- **Process Pipeline:**

    1. **Segmentation:** Lesion segmentation was performed using SAM, with a bounding box prompt encompassing the full image (this specific prompting strategy was noted as a potential issue later).
    2. **LBP Feature Extraction:** A Local Binary Pattern (LBP) histogram was extracted from the grayscale version of the SAM-segmented lesion region to capture texture information.
    3. **Deep Feature Extraction:** The SAM-segmented region was resized to 600x600 pixels. Deep features were then extracted using a pre-trained EfficientNetB7 model (with ImageNet weights, include_top=False to remove the original classification head).
    4. **Feature Concatenation:** The LBP histogram features and the EfficientNetB7 deep features were concatenated to form a combined feature vector for each image.
    5. **SVM Classification:** An SVM classifier was trained on these combined feature vectors derived from the balanced training set (4116 images), using an 80/20 split for training and validation within this set.

3.5.2 Revised Direction: EfficientNetB7-Centered Feature Extraction with ML Classifiers Due to significant performance degradation of the initial hybrid model on the actual test set, the strategy shifted towards simpler pipelines centered on the strong feature extraction capabilities of EfficientNetB7.

-

Approach 1: EfficientNetB7 Features + ML Classifiers

1. **Feature Extraction:**

    - A pre-trained EfficientNetB7 **[17]** model (ImageNet weights, include_top=False) was used. A GlobalAveragePooling2D layer was added after the convolutional base to produce a fixed-size feature vector (2560 dimensions) for each image.
    - Each of the 4116 images from the balanced training set was resized to 600x600 pixels before feature extraction.
    - The extracted features were saved (e.g., as efficientnet_b7_features_approach1.npy) and standardized (scaled) using StandardScaler fitted on these training features.

2. **ML Classification:**

    - The extracted and scaled training features were split into an 80% training subset and a 20% validation subset.
    - Several ML classifiers (SVM with RBF kernel, Logistic Regression, Random Forest) were trained on the 80% feature subset and evaluated on the 20% validation feature subset.
    - **Hyperparameter Tuning:** GridSearchCV (with 3-fold cross-validation) was subsequently used to tune the hyperparameters of SVM (RBF), Logistic Regression, and Random Forest classifiers on the *full set* of scaled EfficientNetB7 training features. The best-tuned models were saved.
    - For test set evaluation, test images were identically preprocessed (resized to 600x600), features extracted, and then scaled using the *scaler fitted on the training data* before prediction with the tuned models.

Approach 2: Combined LBP + EfficientNetB7 Features + ML Classifiers

•

1. **Feature Extraction:**

    - For each of the 4116 balanced training images:

- LBP features (radius 3, 24 points, 'uniform' method) were extracted from the original size grayscale version of the preprocessed (hair removed, denoised, enhanced) image.
- EfficientNetB7 features were extracted from the 600x600 resized color version of the preprocessed image (as in Approach 1).
- These two feature sets were concatenated, resulting in a 2586-dimensional feature vector per image, which was saved (e.g., combined_lbp_effnet_features_approach2.npy) and scaled.

2. **ML Classification:** The same process of 80/20 validation split evaluation and subsequent hyperparameter tuning (GridSearchCV) on the full combined training features was followed as in Approach 1. The best-tuned models were saved, and test set evaluation followed the same protocol.

3.6 Evaluation Strategy

To assess the performance of the various segmentation and classification methods, a set of standard evaluation metrics was employed.

- **For Segmentation Models (MSER, U-Net, SAM):**
  - **Dice Coefficient (Sørensen-Dice Index):** Measures the overlap between the predicted segmentation mask and the ground truth mask. Dice= $|X| + |Y| 2\times |X\cap Y|$, where X is the prediction and Y is the ground truth.
  - **Intersection over Union (IoU / Jaccard Index):** Also measures overlap. IoU= $|X\cup Y|$ $|X\cap Y|$.
  - **Pixel Accuracy:** The percentage of pixels correctly classified (lesion or background). These metrics were calculated by comparing the model's output masks against the ISIC-2017_Test_v2_Part1_GroundTruth.

- **For Classification Models (SVM, Random Forest, EfficientNetB7-based):**
    - **Accuracy:** Overall correct classification rate.
    - **Precision, Recall (Sensitivity), F1-Score:** Calculated per class to understand performance on individual categories, especially important for the imbalanced Melanoma and SK classes.
    - **Confusion Matrix:** To visualize the distribution of true vs. predicted classes.

3.7 Chapter Summary

This chapter has provided a detailed account of the materials and the multi-phase methodology adopted for this thesis. We began by describing the ISIC 2017 dataset and the computational environment. The extensive data preprocessing pipeline, including hair removal, denoising, enhancement, resizing, and a crucial data augmentation strategy to balance the training classes, was then detailed. Following this, the chapter outlined the experimental approaches for lesion segmentation, covering MSER, U-Net, a hybrid MSER+U-Net, and the SAM foundation model. Subsequently, the methodologies for lesion classification were presented, starting from an initial complex hybrid model and evolving to strategies centered around EfficientNetB7 as a feature extractor with various ML classifiers, and culminating in the ongoing work on an end-to-end fine-tuning of an attention-enhanced EfficientNetB7. Finally, the evaluation metrics and dataset splitting strategies used to assess model performance at different stages were summarized. This systematic approach, with its iterative refinements based on observed outcomes, forms the core of the investigative work presented in this thesis. The results stemming from these methodologies will be presented and analyzed in the subsequent chapter.

Chapter 4: Experimental Results and Discussion

4.1 Introduction:

This chapter presents a comprehensive account and critical analysis of the experimental results obtained from the methodologies detailed in Chapter 3. The primary objective of this research was to investigate and develop effective automated methods for skin lesion segmentation and classification using the ISIC 2017 dataset. Accordingly, this chapter is structured to first present the outcomes of the various segmentation techniques explored, followed by an in-depth exposition of the results from the multi-stage classification experiments.

The classification experiments progressed from an initial hybrid approach leveraging segmented regions to more refined strategies utilizing whole-image features with advanced deep learning models like EfficientNetB7 and various machine learning classifiers. The results of an attempted end-to-end fine-tuning of an attention-enhanced EfficientNetB7 model are also reported. For each experimental phase, quantitative results will be presented through tables and figures, including confusion matrices and Receiver Operating Characteristic (ROC) curves. Qualitative visual examples will be provided for segmentation tasks.

Crucially, this chapter will also feature a detailed discussion section. This discussion will compare the performance of different methods, analyze the significant discrepancy observed between validation and test set performances, and delve into the underlying reasons such as domain shift, class imbalance, and the impact of data augmentation. Computational challenges encountered during the research will also be addressed, leading to an understanding of the study's limitations and key takeaways.

4.2 Experimental Setup Recap

Before presenting the results, it is pertinent to briefly reiterate the key components of the experimental setup, elaborated in Chapter 3, which form the context for the findings discussed herein.

- **Computational Environment:**

- **Local Machine:** ASUS laptop with Intel® Core™ i5-8250U CPU @ 1.60GHz, 8.00 GB RAM, NVIDIA GeForce GTX 1050 & Intel® UHD Graphics 620, 64-bit Windows 10 Pro OS. This was used for initial development and some experiments.
- **Google Colab:** Utilized for more computationally intensive tasks, particularly deep learning model training and fine-tuning, providing access to a T4 GPU environment.(was used in the **Attention-Enhanced EfficientNetB7** only**)**

•

Key Software and Libraries:

- **Python** served as the primary programming language.(3.12.6)
- **OpenCV** was used for image processing operations.
- **TensorFlow/Keras** were employed for deep learning model implementation (U-Net, EfficientNetB7).
- **PyTorch** and the segment-anything library were used for the Segment Anything Model (SAM).
- **Scikit-learn** was utilized for machine learning classifiers (SVM, Random Forest, etc.), standard evaluation metrics, feature scaling (StandardScaler), data splitting (train_test_split), and hyperparameter optimization (GridSearchCV).
- **DEAP** was noted for previous Genetic Algorithm-based hyperparameter optimization.
- **Pandas**, **NumPy**, **Matplotlib/Seaborn**, and **Joblib** were used for data handling, numerical computation, plotting, and model persistence, respectively.

4.3 Segmentation Experiments and Results

Accurate segmentation of the lesion from the surrounding skin is a critical upstream task in many CAD systems, as it isolates the region of interest for subsequent feature extraction and analysis. This section details the comparative performance of the segmentation methods evaluated in this study.

Introduction to Segmentation Task: The goal of the segmentation task was to produce a binary mask that accurately delineates the boundaries of the skin lesion in dermoscopic images.

Methods Evaluated: As described in Chapter 3.4, the following methods were evaluated:

- Maximally Stable Extremal Regions (MSER) with default parameters.
- U-Net, a deep learning model specifically trained for this task.
- MSER + U-Net, a hybrid approach.
- Segment Anything Model (SAM) with a ViT-B backbone, used in a zero-shot setting with manual bounding box prompts.

Quantitative Segmentation Results: The performance of these methods was evaluated on the ISIC 2017 Test Set (600 images with ground truth masks). The results are summarized in Table 4.1.

Table 2 : Comparative Performance of Segmentation Methods on ISIC 2017 Test Set

Model	Dice Score	IoU Score	Accuracy
Default MSER	0.0081	0.0046	0.7541
MSER + U-Net	0.6133	0.5250	0.8701
U-Net	0.7697	0.6796	0.9064
SAM (ViT-B, Manual Prompts)	0.7887	0.6718	0.9072

Qualitative Segmentation Results (Visual Examples): To complement the quantitative metrics, visual examples of the segmentation outputs for each method are crucial.

Default Mser:

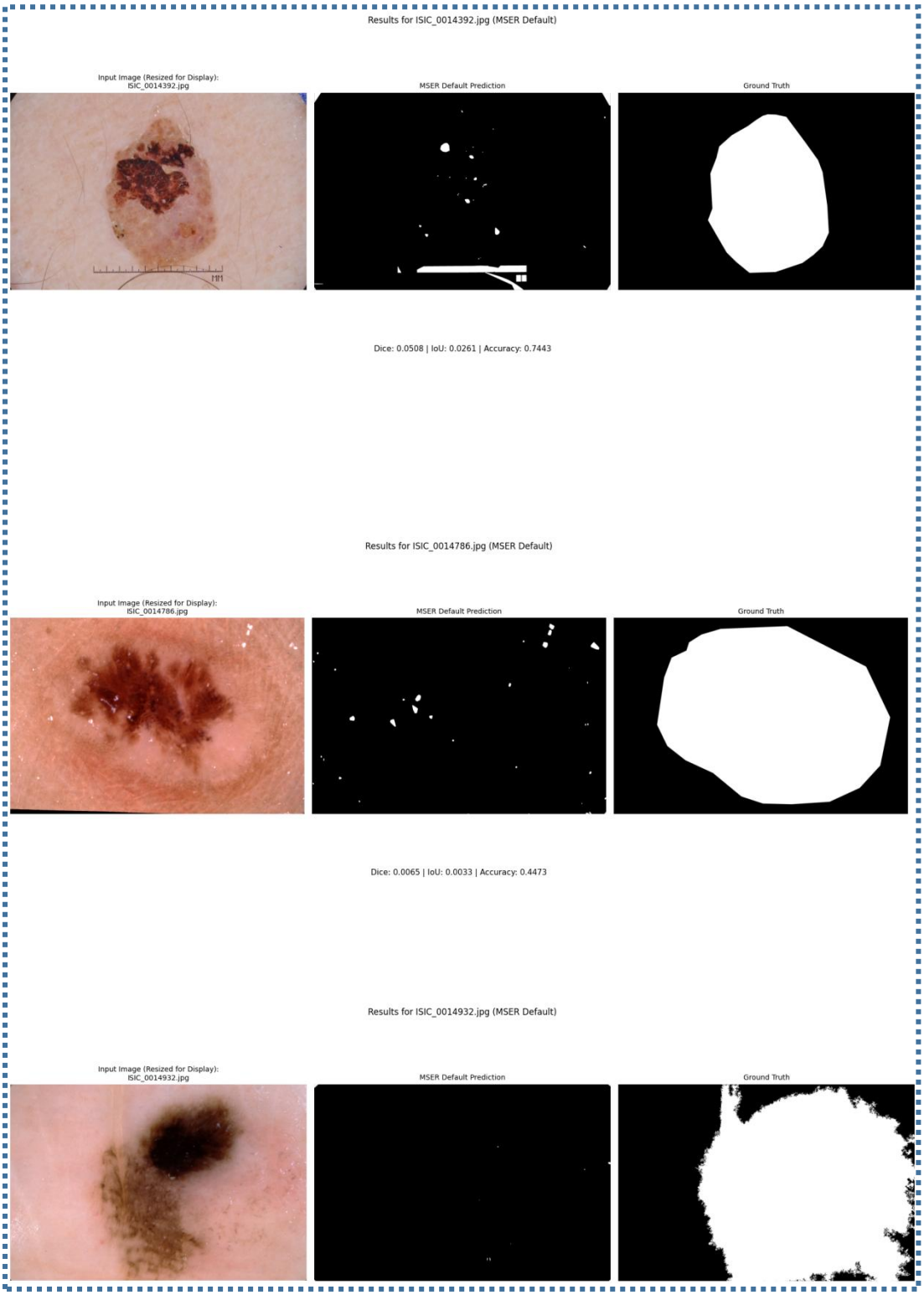


Figure 8 : demonstration of Mser (Default) Segmentation results example on 3 images from the test set

UNET:

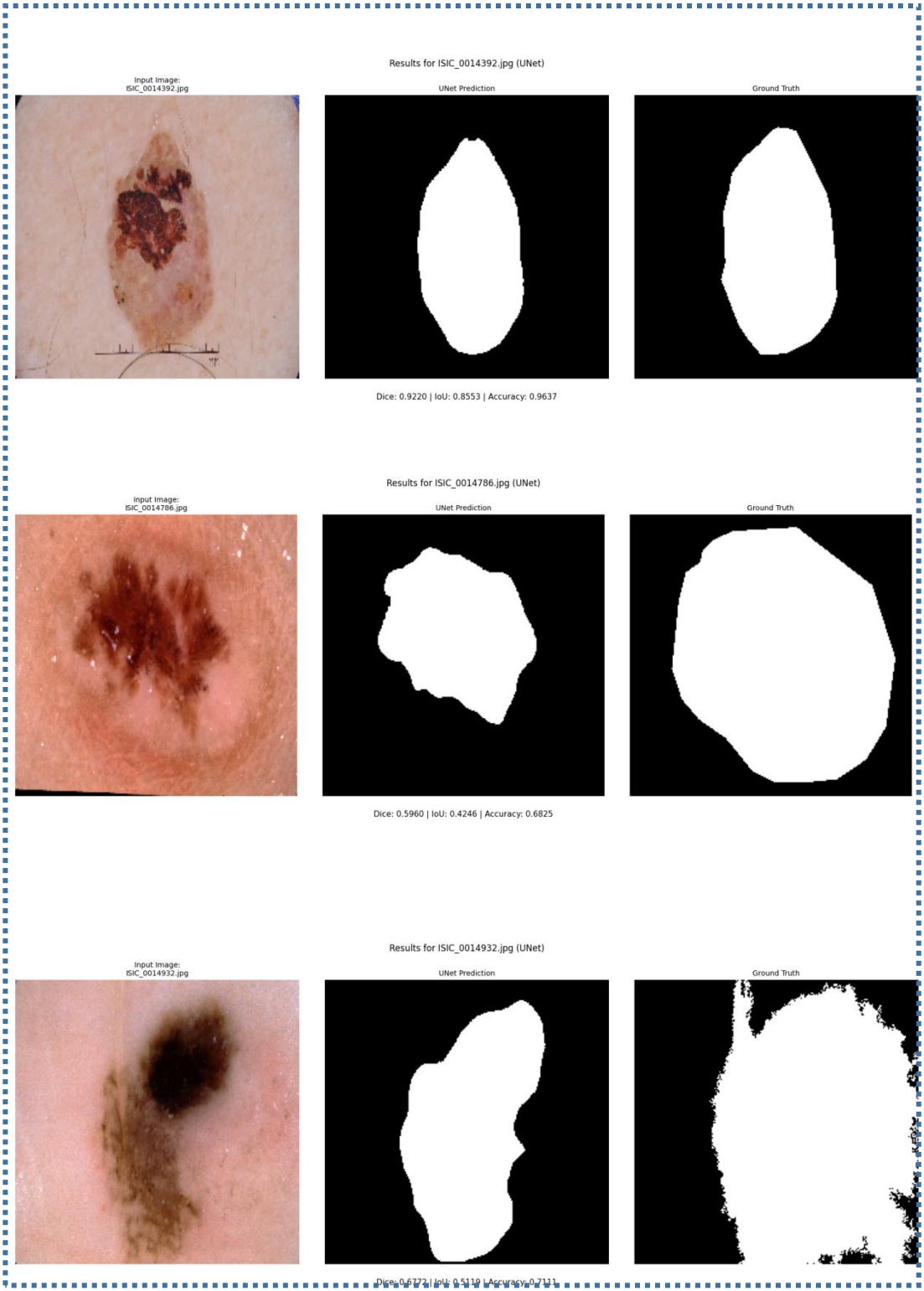


Figure 9 : demonstration of UNet Segmentation results example on 3 images from the test set

Mser+UNet:

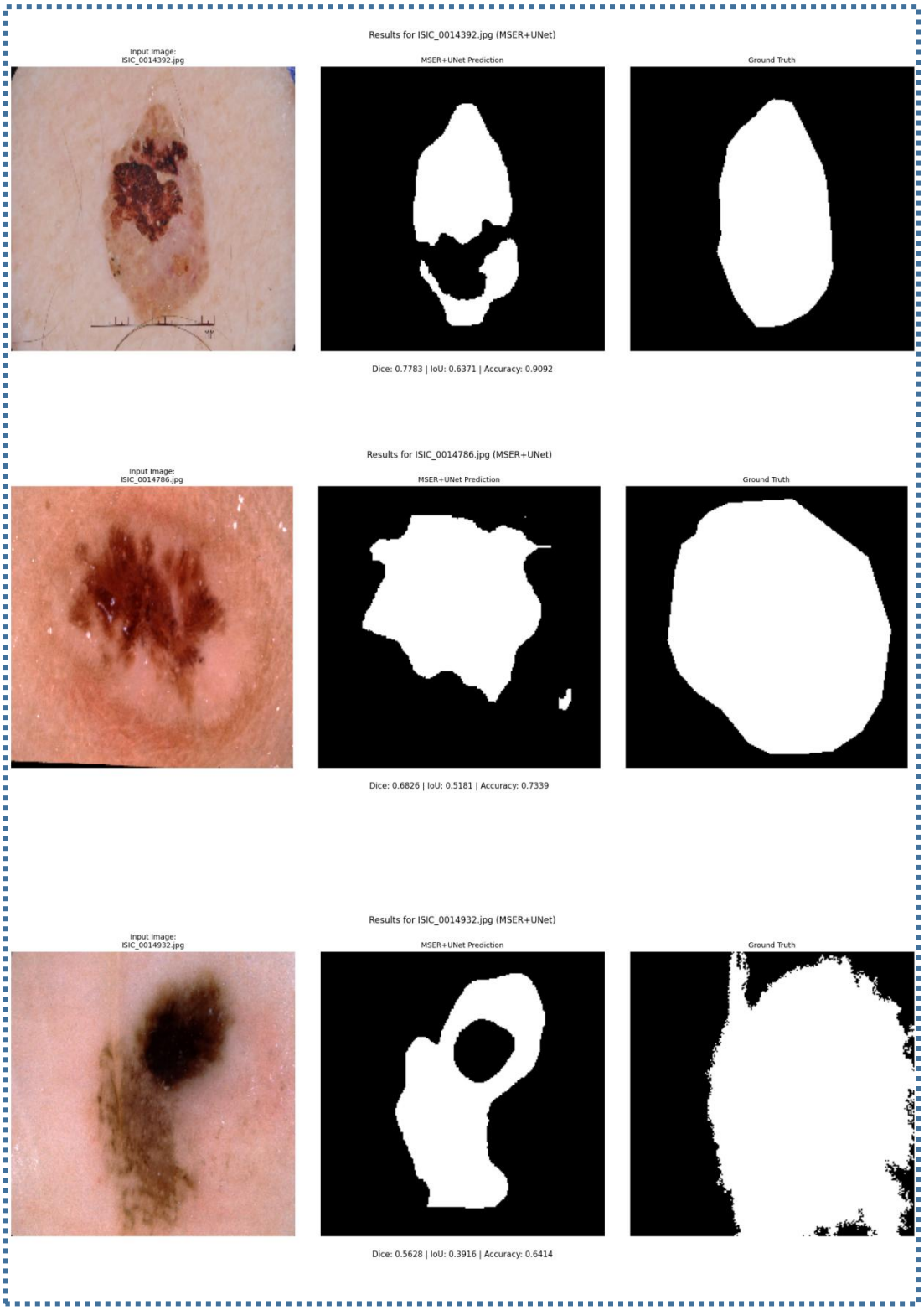


Figure 10 : demonstration of Mser+UNet Segmentation results example on 3 images from the test set

SAM:

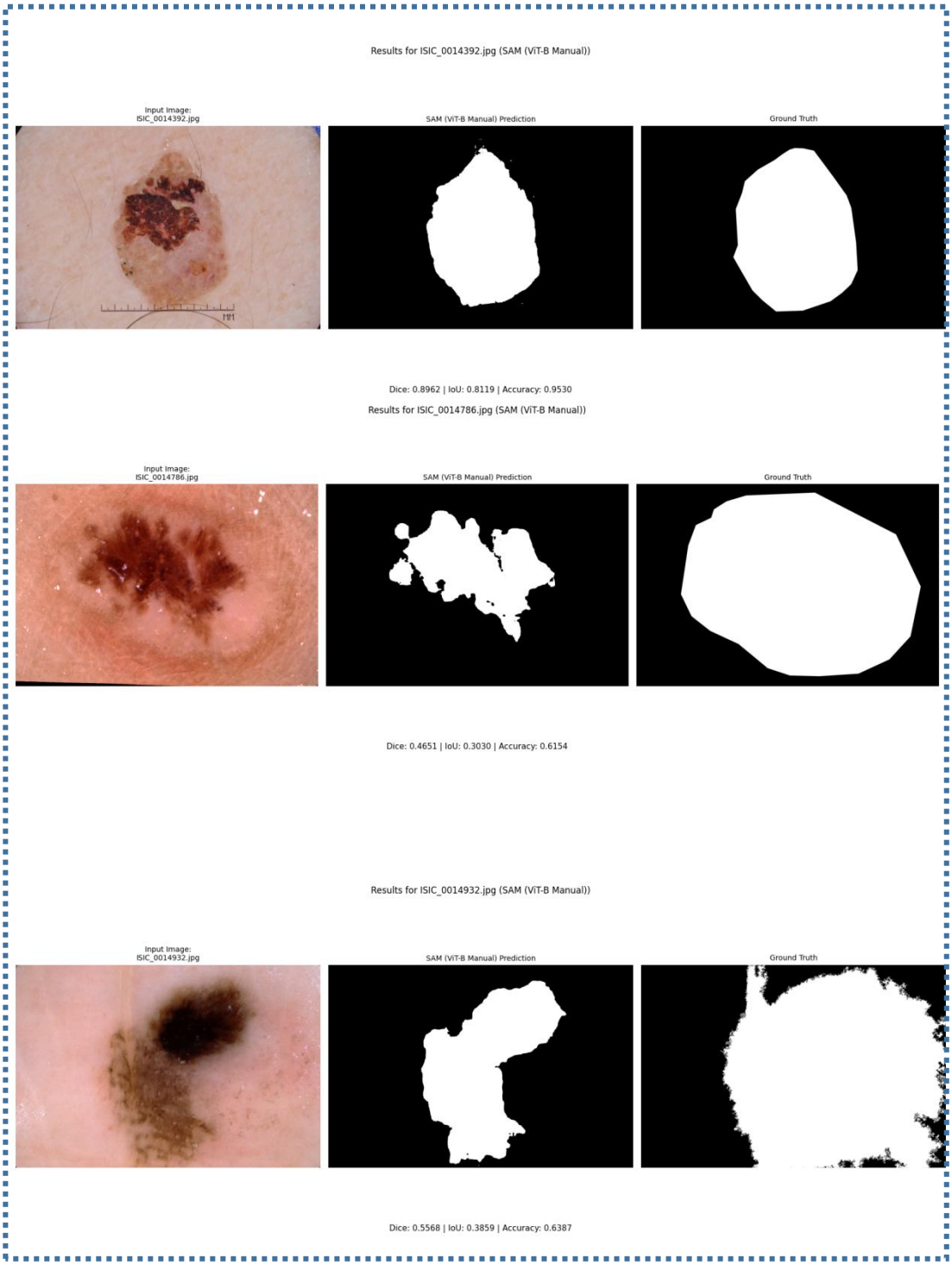


Figure 11 : demonstration of SAM Segmentation results example on 3 images from the test set

4.3.5 Brief Discussion of Segmentation Results:

The results from the segmentation experiments reveal several key insights:

- **Default MSER:** This traditional computer vision algorithm performed extremely poorly, with Dice and IoU scores near zero. The relatively high pixel accuracy (0.7541) is misleading, primarily due to the model correctly classifying the large background areas, while failing to identify the lesion itself. This underscores the unsuitability of default MSER for this complex task without extensive parameter tuning.
- **U-Net:** The U-Net model, trained specifically on the ISIC 2017 dataset, demonstrated strong performance, achieving a Dice score of 0.7697 and an IoU of 0.6796. This highlights the effectiveness of deep learning architectures with their learned hierarchical features for robust lesion localization in biomedical images.
- **MSER + U-Net:** The hybrid MSER+U-Net approach did not improve upon the standard U-Net and, in fact, performed considerably worse (Dice: 0.6133). This suggests that the MSER-derived masks, even when using optimized parameters (as noted in Chapter 3), might have introduced noise or conflicting information that hindered the U-Net's learning process, or the simple concatenation fusion strategy was suboptimal.
- **SAM (ViT-B with Manual Prompts):** The Segment Anything Model, used in a zero-shot setting with manual bounding box prompts, achieved the highest Dice score (0.7887) and accuracy (0.9072) among the evaluated methods. This showcases the remarkable generalization capabilities of foundation models, even without task-specific fine-tuning. However, its performance is dependent on the quality and placement of the manual prompt. The results are promising, especially considering other research like "SkinSAM" has shown further improvements with fine-tuning.
- **Metric Importance:** The results reaffirm that Dice and IoU are more insightful metrics than pixel accuracy for imbalanced segmentation tasks like lesion segmentation, where the foreground (lesion) is often much smaller than the background.

In conclusion, for direct lesion segmentation, deep learning-based methods (U-Net and SAM) significantly outperformed the traditional MSER algorithm. SAM, even without fine-tuning, showed state-of-the-art potential when appropriately prompted.

4.4 Classification Experiments and Results

This section details the results of the various classification pipelines developed to categorize skin lesions into Benign (Nevus), Melanoma, or Seborrheic Keratosis (SK). All classification experiments utilized the balanced training dataset of 4116 preprocessed and augmented images. An 80/20 stratified split (3292 training, 824 validation samples) was used for initial model evaluation during development and for certain stages of hyperparameter tuning. The final, definitive evaluation was performed on the independent, imbalanced ISIC 2017 Test Set (600 images).

4.4.1 Initial Hybrid Approach (SAM-Segmented Region Features - "Algorithm 1") : This first approach, inspired by Algorithm 1 from the supervisor's guidance, involved segmenting the lesion using SAM (with a full image bounding box prompt), then extracting LBP texture features from the grayscale segmented region and deep features using EfficientNetB7 from the color segmented region (resized to 600x600). These concatenated features were then classified using an SVM with an RBF kernel.

Results on Validation Split (80/20 of balanced training data):

- Overall Accuracy: **~81.19%**
- Detailed Classification Report (Validation Split):

Table 3 : Classification Performance of Hybrid Approach (SAM+LBP+EffNetB7+SVM) on Validation Split

Class	Precision	Recall	F1-score	Support
Benign (0)	0.75	0.97	0.84	274
Melanoma (1)	0.90	0.64	0.75	275
SK (2)	0.84	0.83	0.83	275

Results on Actual ISIC 2017 Test Set (600 images):

- Overall Accuracy: **~66.33%**
- Detailed Classification Report (Actual Test Set):

Table 4 : Classification Performance of Hybrid Approach (SAM+LBP+EffNetB7+SVM) on ISIC 2017 Test Set

Class	Precision	Recall	F1-score	Support
Benign (0)	0.68	0.96	0.79	393
Melanoma (1)	0.00	0.00	0.00	117
SK (2)	0.57	0.22	0.32	90

The drastic drop in performance, particularly the complete failure to recall Melanoma cases (Recall = 0.00) on the test set, highlighted significant issues with this initial hybrid approach, primarily attributed to domain shift and the model's inability to generalize from the balanced, augmented training distribution (and SAM-segmented features) to the imbalanced, real-world test set.

4.4.2 Revised Approaches (Whole Image Features & ML Classifiers):

Given the challenges with the segmentation-based hybrid model, the strategy shifted to using features extracted from the entire preprocessed image, leveraging the power of EfficientNetB7 as a feature extractor.

4.4.2.1 Approach 1: EfficientNetB7 Features Only In this approach, each of the 4116 balanced training images was resized to 600x600, preprocessed using *efficientnet.preprocess_input*, and fed into a pre-trained EfficientNetB7 (ImageNet weights, *include_top=False*) followed by a *GlobalAveragePooling2D* layer to extract 2560-dimensional feature vectors.

- **ML Classification (Default Parameters on Validation Split - 80/20 of training features):** The extracted features were split, scaled using `StandardScaler`, and used to train various ML classifiers.

Table 5 : Performance of ML Classifiers with EfficientNetB7 Features (Validation Split)

Classifier	Accuracy	Benign (P/R/F1)	Melanoma (P/R/F1)	SK (P/R/F1)
SVM (RBF, C=1)	0.8410	0.77/0.95/0.85	0.92/0.67/0.77	0.88/0.91/0.89
SVM (Linear, C=1)	0.8216	0.79/0.83/0.81	0.83/0.75/0.79	0.85/0.89/0.87
Random Forest (200 trees)	0.8070	0.75/0.97/0.85	0.87/0.63/0.73	0.84/0.83/0.83
Logistic Regression	0.8301	0.81/0.83/0.82	0.82/0.75/0.78	0.86/0.91/0.88
Gradient Boosting (100 trees)	0.8155	0.78/0.93/0.85	0.83/0.68/0.75	0.84/0.84/0.84
KNN (k=7)	0.8107	0.76/0.92/0.83	0.81/0.70/0.75	0.88/0.82/0.85

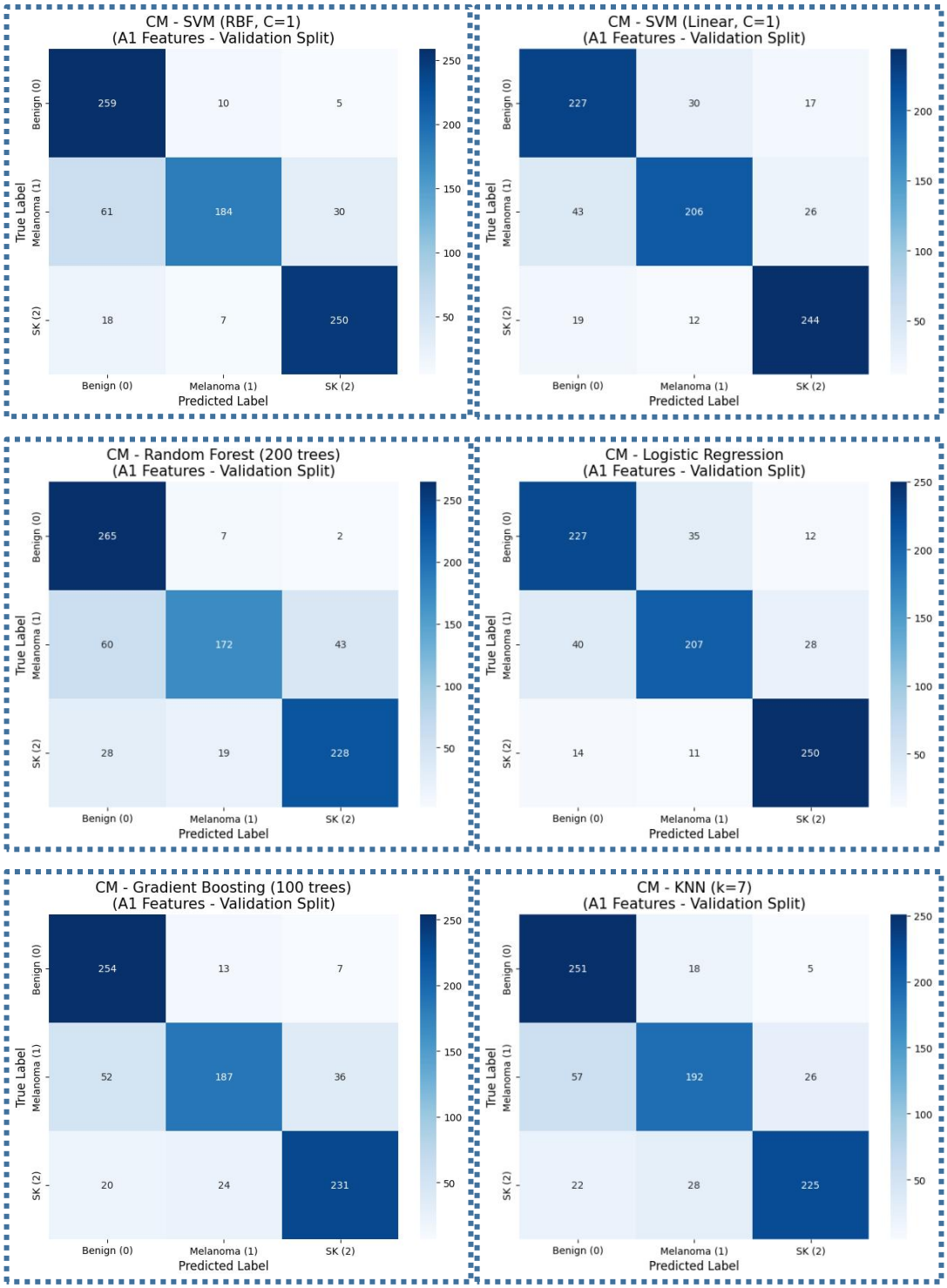
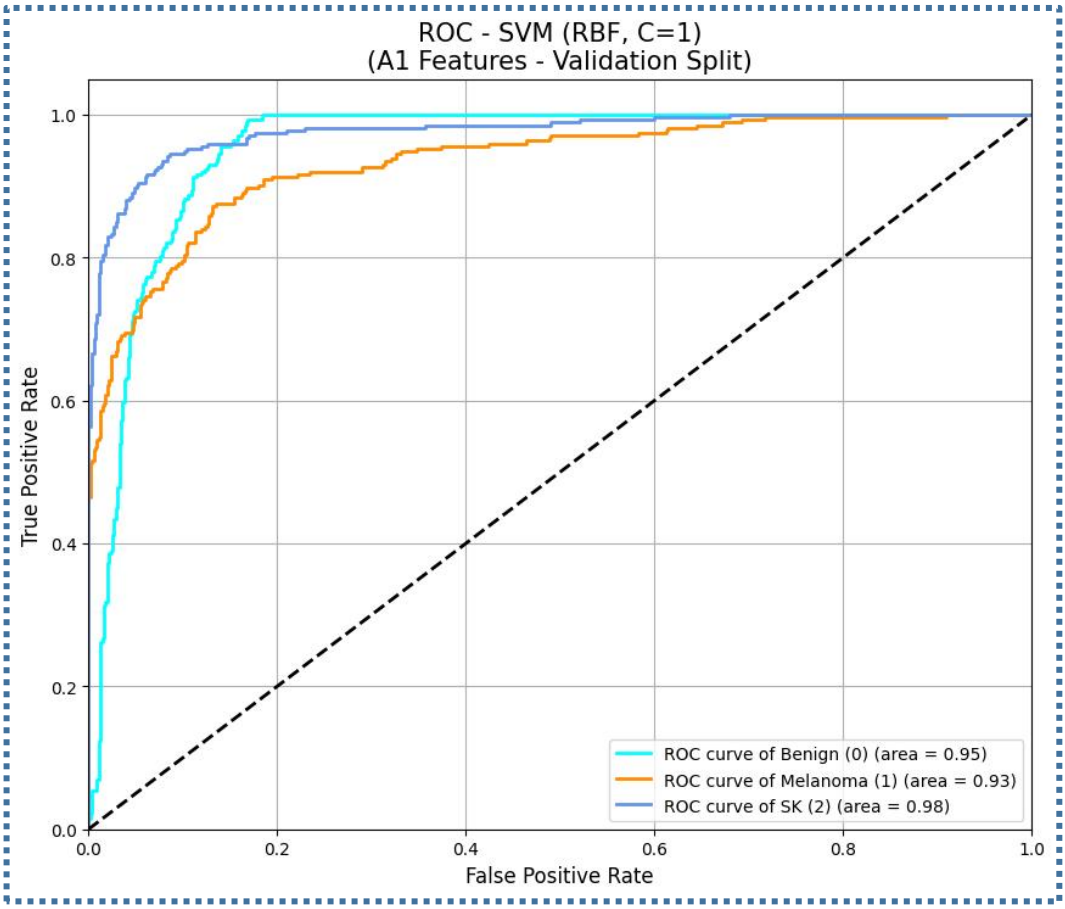
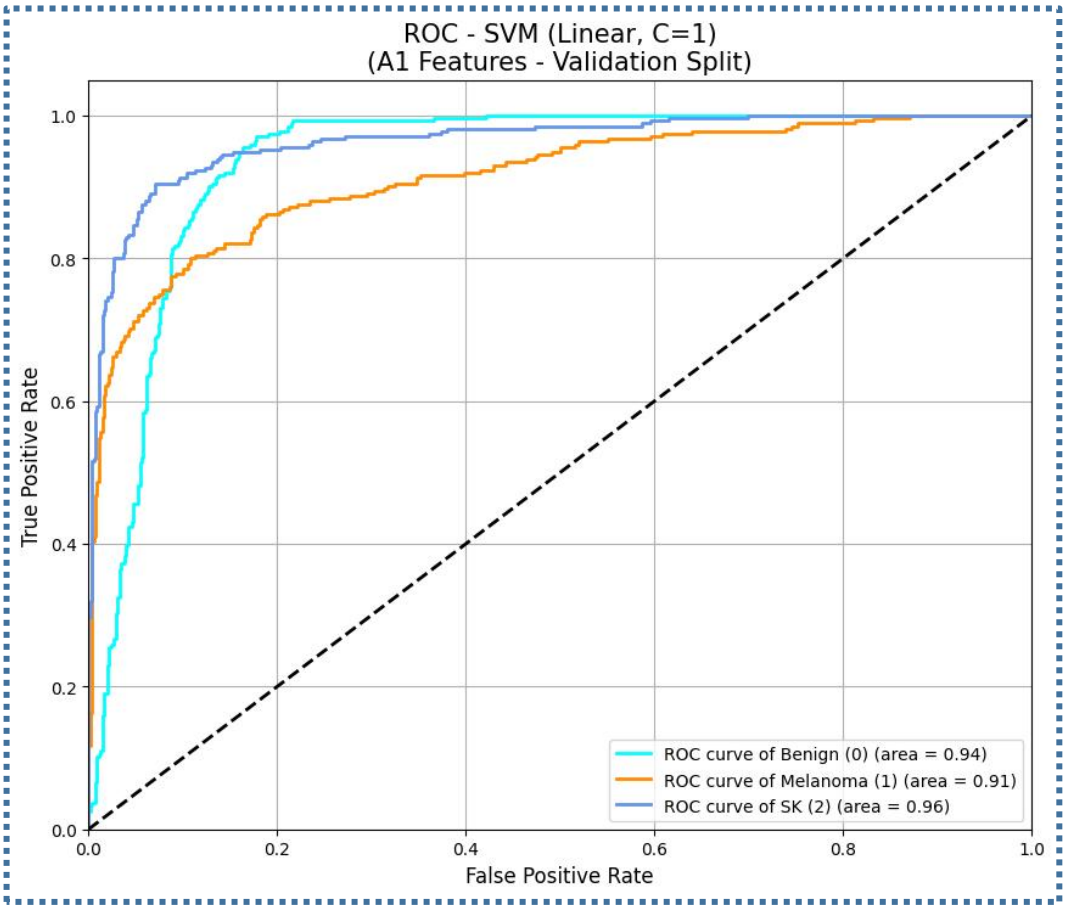
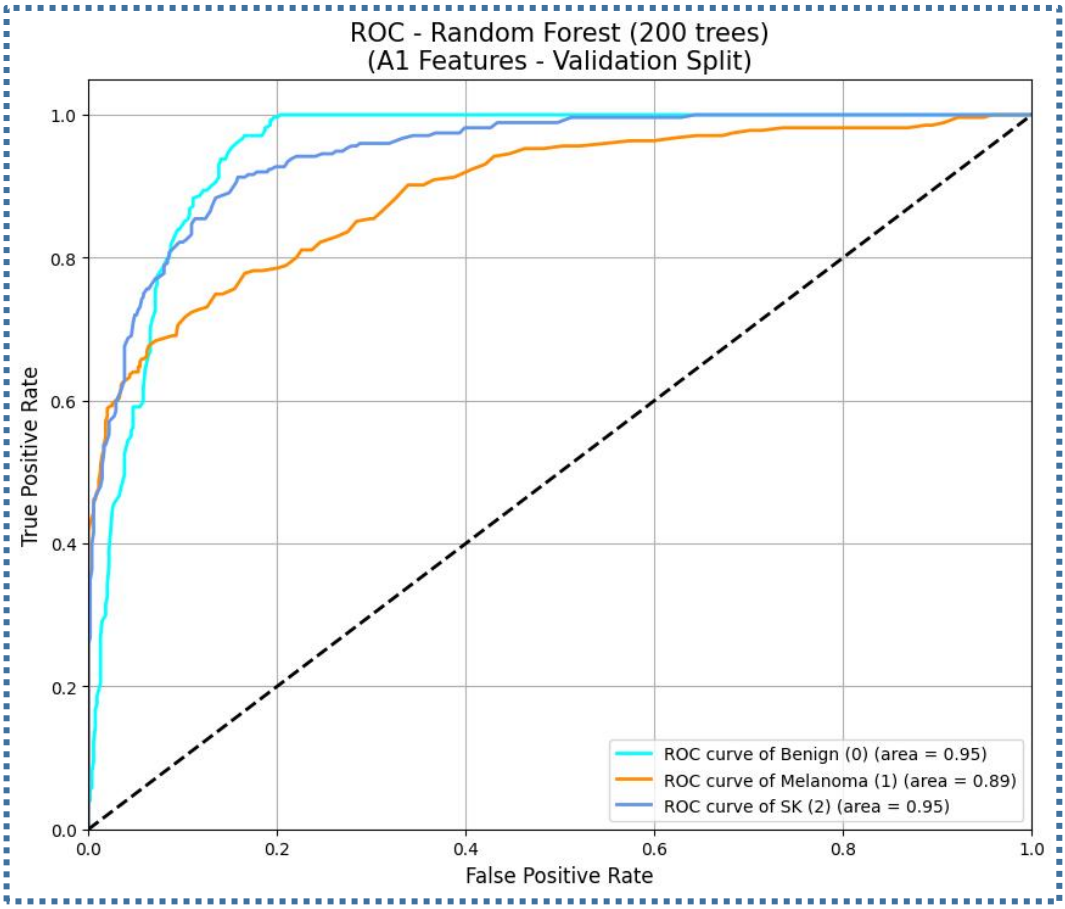
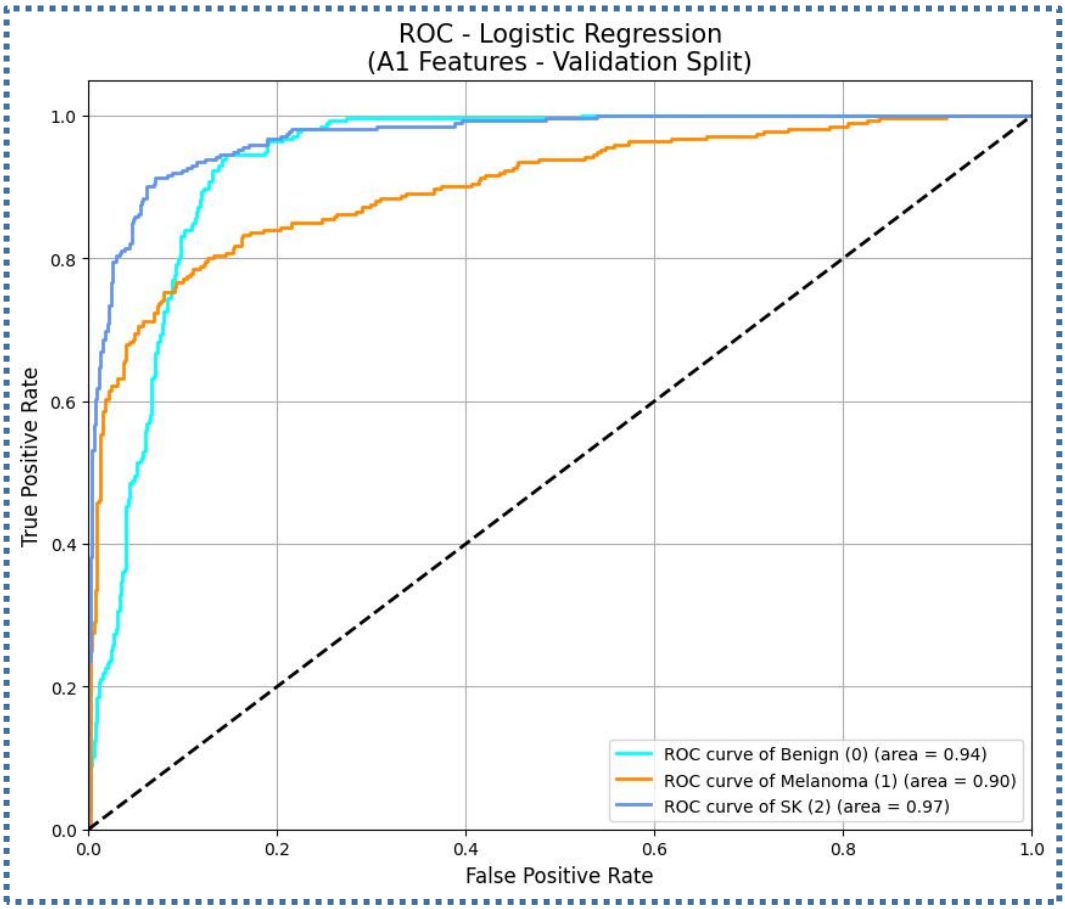


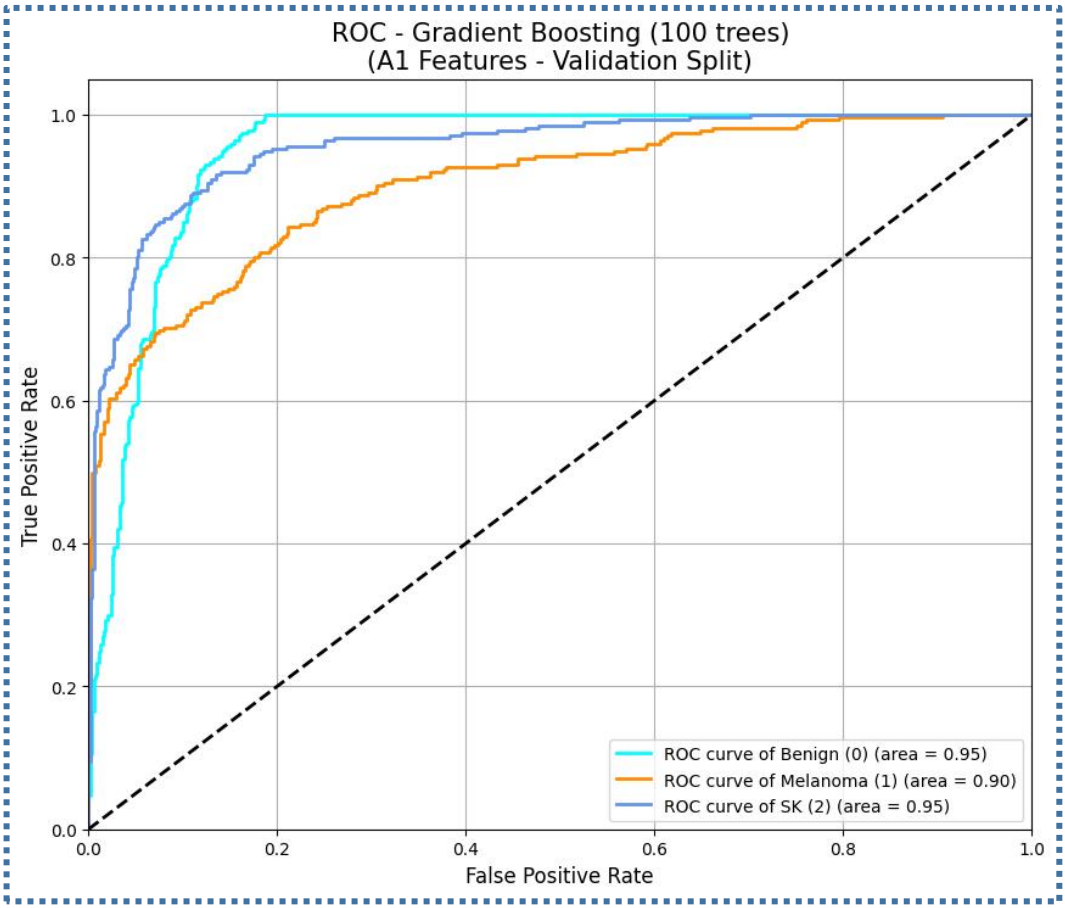
Figure 12 : Confusion Matrices for the ML Classifiers with EfficientNetB7 Features (Validation Split)

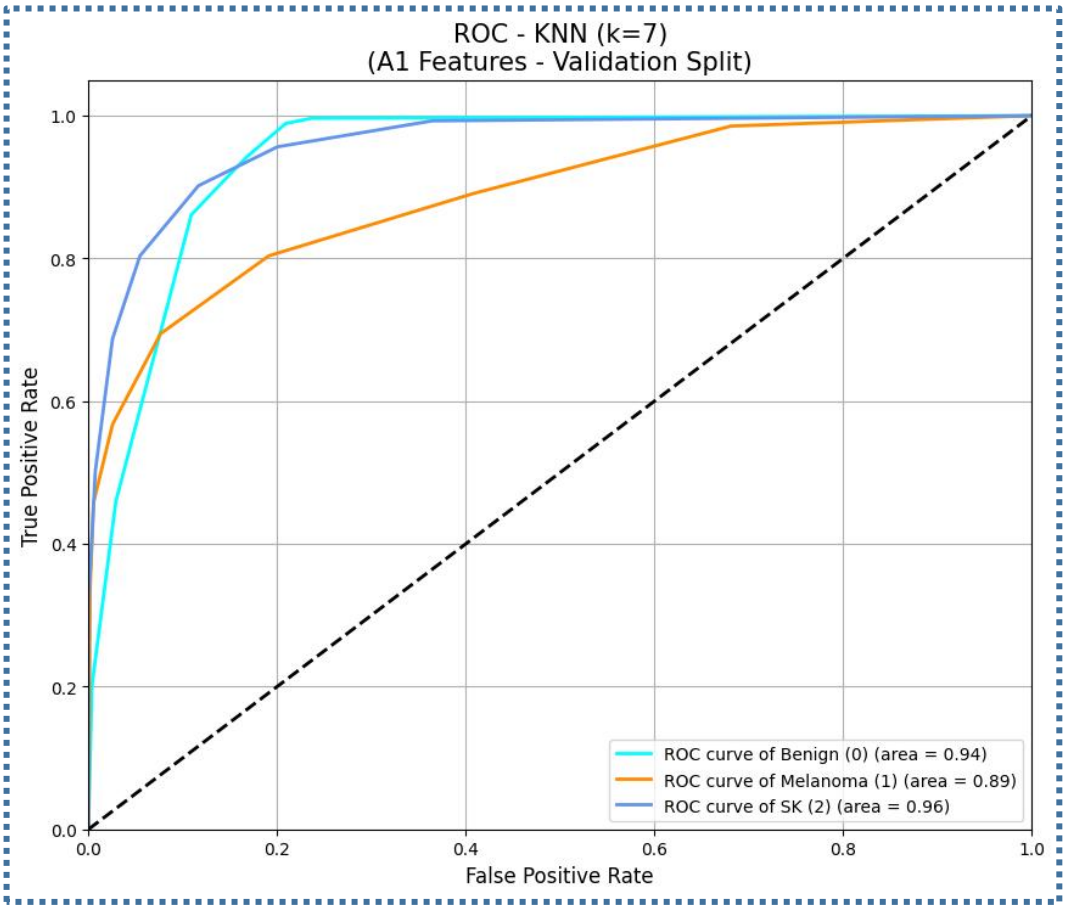

ROC - SVM (RBF, C=1)
(A1 Features - Validation Split)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of Benign (0) (area = 0.95)
ROC curve of Melanoma (1) (area = 0.93)
ROC curve of SK (2) (area = 0.98)


ROC - SVM (Linear, C=1)
(A1 Features - Validation Split)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of Benign (0) (area = 0.94)
ROC curve of Melanoma (1) (area = 0.91)
ROC curve of SK (2) (area = 0.96)


ROC - Random Forest (200 trees)
(A1 Features - Validation Split)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of Benign (0) (area = 0.95)
ROC curve of Melanoma (1) (area = 0.89)
ROC curve of SK (2) (area = 0.95)


ROC - Logistic Regression
(A1 Features - Validation Split)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of Benign (0) (area = 0.94)
ROC curve of Melanoma (1) (area = 0.90)
ROC curve of SK (2) (area = 0.97)


ROC - Gradient Boosting (100 trees)
(A1 Features - Validation Split)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of Benign (0) (area = 0.95)
ROC curve of Melanoma (1) (area = 0.90)
ROC curve of SK (2) (area = 0.95)


ROC - KNN (k=7)
(A1 Features - Validation Split)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of Benign (0) (area = 0.94)
ROC curve of Melanoma (1) (area = 0.89)
ROC curve of SK (2) (area = 0.96)

Hyperparameter Tuning: GridSearchCV (3-fold CV) was used on the full set of 4116 scaled EfficientNetB7 training features to tune SVM RBF, Logistic Regression, and Random Forest classifiers. Genetic Algorithms (GA) were also explored for SVM tuning in earlier stages.

- **Evaluation of Tuned Models on Actual ISIC 2017 Test Set (600 images):** Test set images were processed to extract EfficientNetB7 features and scaled using the scaler fitted on the full training data.

Table 6 : Performance of Tuned ML Classifiers with EfficientNetB7 Features (Actual Test Set)

Tuned Classifier	Accuracy	Benign (P/R/F1)	Melanoma (P/R/F1)	SK (P/R/F1)
Random Forest	0.6567	0.69/0.95/0.80	0.33/0.03/0.06	0.34/0.17/0.22
Logistic Regression	0.6433	0.70/0.93/0.80	0.20/0.02/0.03	0.27/0.19/0.22
SVM (RBF)	0.6383	0.69/0.91/0.79	0.42/0.04/0.08	0.29/0.23/0.26

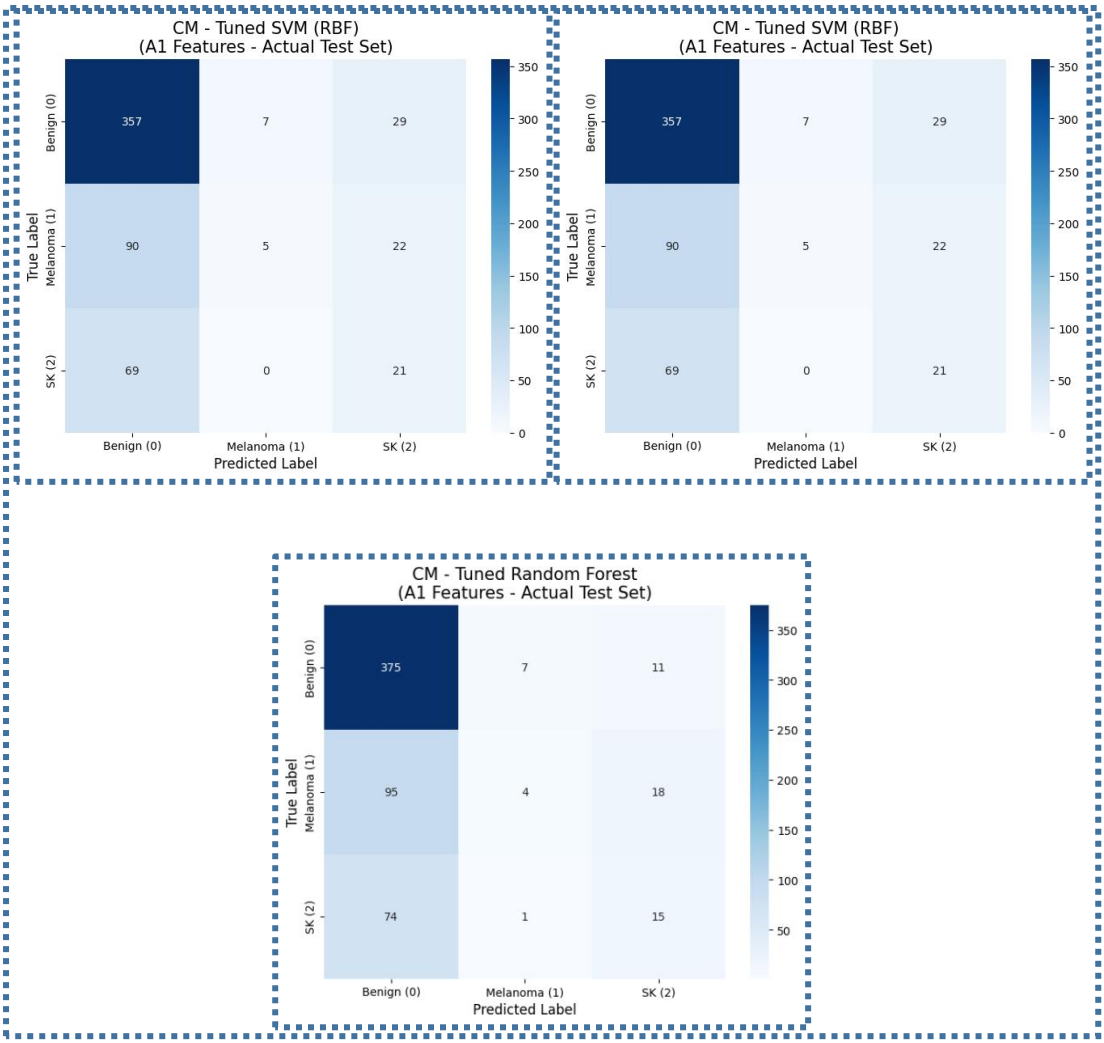
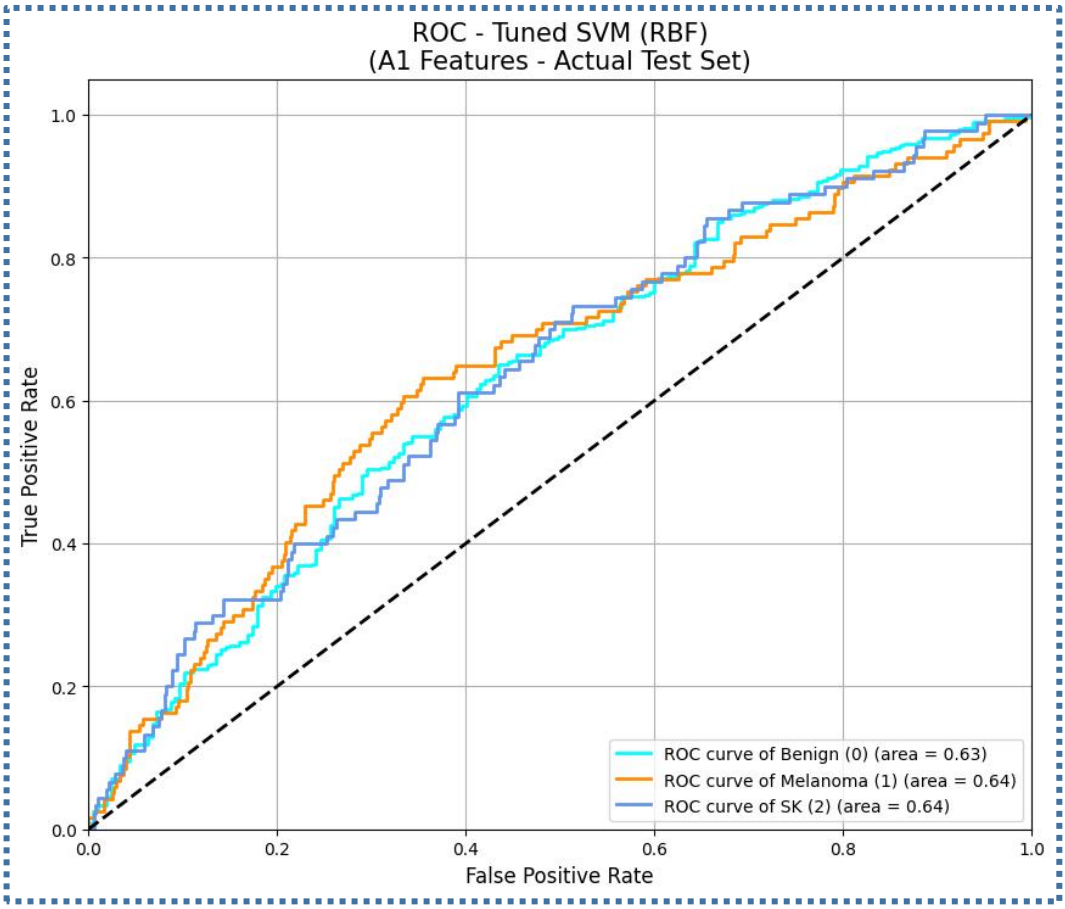
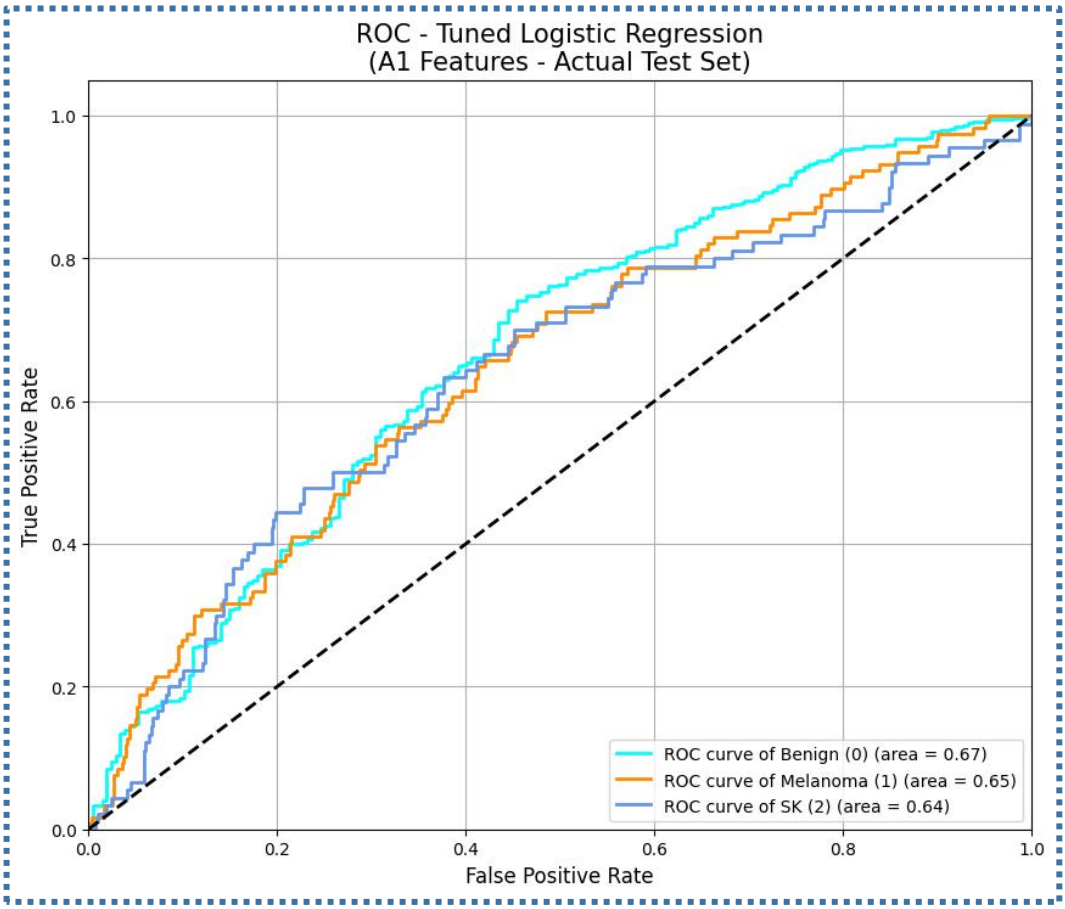
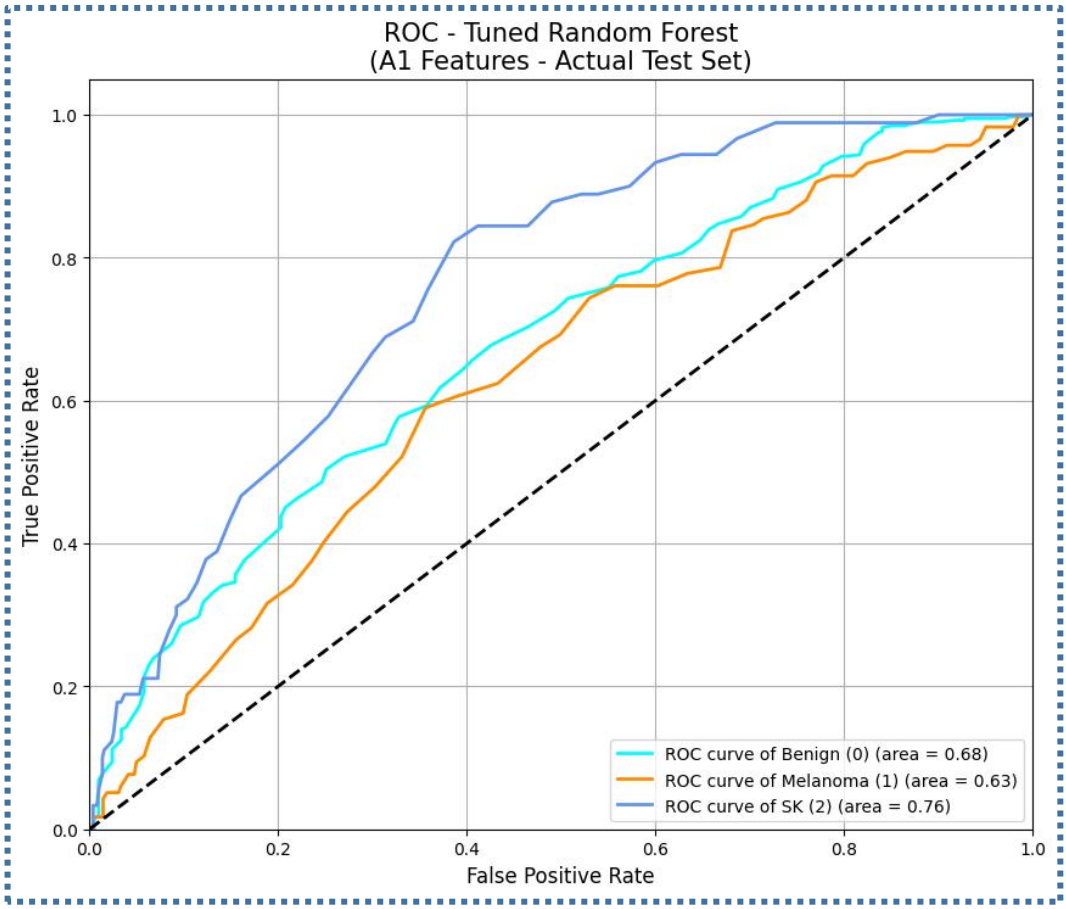


Figure 13 : Confusion Matrices for the Tuned ML Classifiers with EfficientNetB7 Features (Actual Test Set)


ROC - Tuned SVM (RBF)
(A1 Features - Actual Test Set)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of Benign (0) (area = 0.63)
ROC curve of Melanoma (1) (area = 0.64)
ROC curve of SK (2) (area = 0.64)


ROC - Tuned Logistic Regression
(A1 Features - Actual Test Set)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of Benign (0) (area = 0.67)
ROC curve of Melanoma (1) (area = 0.65)
ROC curve of SK (2) (area = 0.64)



4.4.2.2 Approach 2: Combined LBP + EfficientNetB7 Features In this approach, LBP features (radius 3, 24 points, 'uniform' method from original size grayscale image) were concatenated with EfficientNetB7 features (from 600x600 color image), resulting in 2586-dimensional feature vectors for each of the 4116 training images.

ML Classification (Default Parameters on Validation Split - 80/20 of training features):

Table 7 : Performance of ML Classifiers with Combined LBP + EfficientNetB7 Features (Validation Split)

Classifier	Accuracy	Benign (P/R/F1)	Melanoma (P/R/F1)	SK (P/R/F1)
SVM (RBF, C=1)	0.8252	0.75/0.95/0.84	0.93/0.64/0.76	0.85/0.88/0.87
SVM (Linear, C=1)	0.7961	0.78/0.81/0.79	0.76/0.72/0.74	0.85/0.86/0.85
Random Forest (200 trees)	0.7937	0.71/0.97/0.82	0.92/0.60/0.72	0.82/0.82/0.82
Logistic Regression	0.8022	0.79/0.80/0.79	0.78/0.73/0.76	0.84/0.87/0.86
Gradient Boosting (100 trees)	0.8034	0.76/0.92/0.83	0.83/0.65/0.73	0.83/0.84/0.83
KNN (k=7)	0.7888	0.74/0.88/0.80	0.80/0.65/0.72	0.84/0.84/0.84

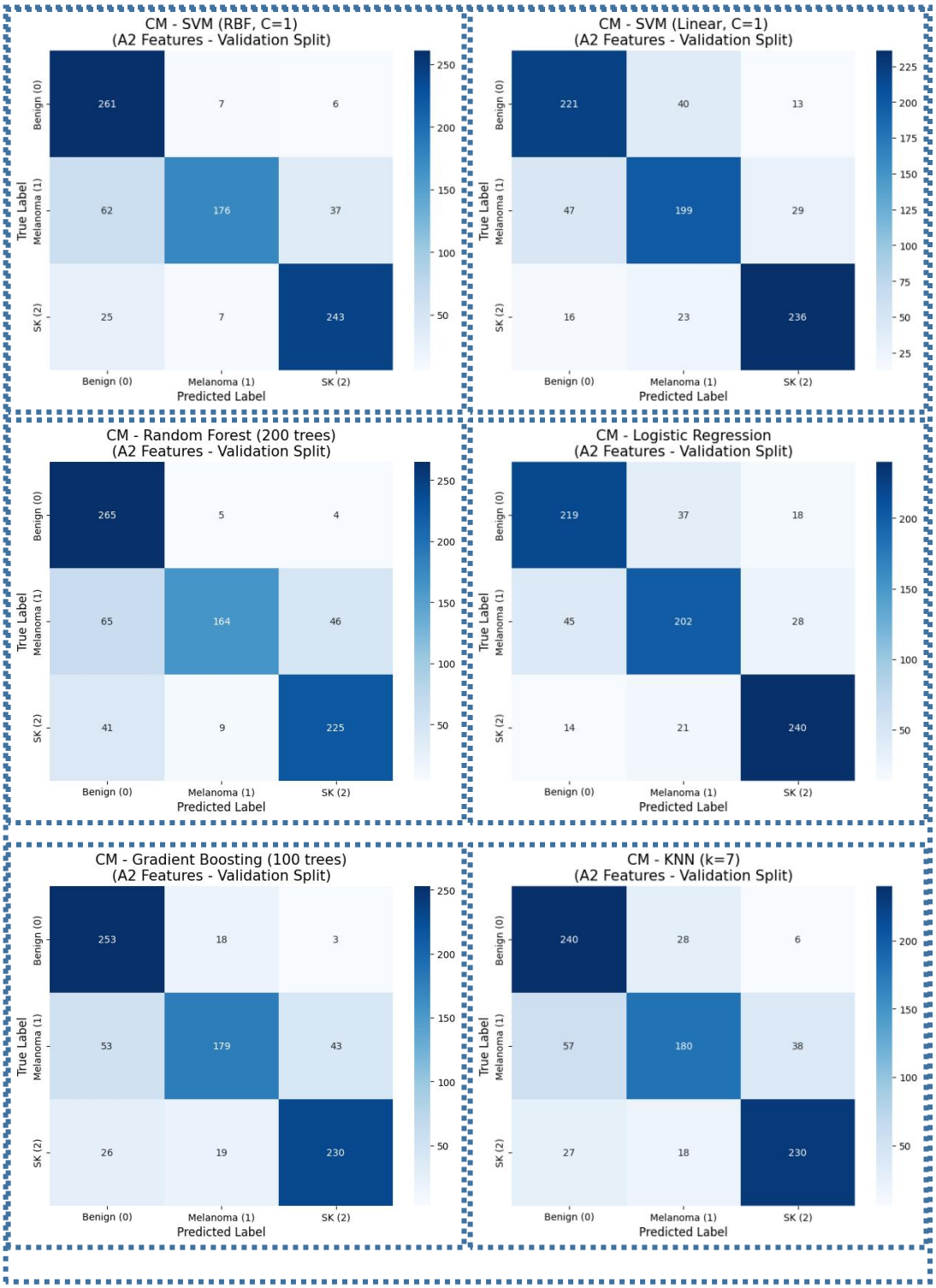
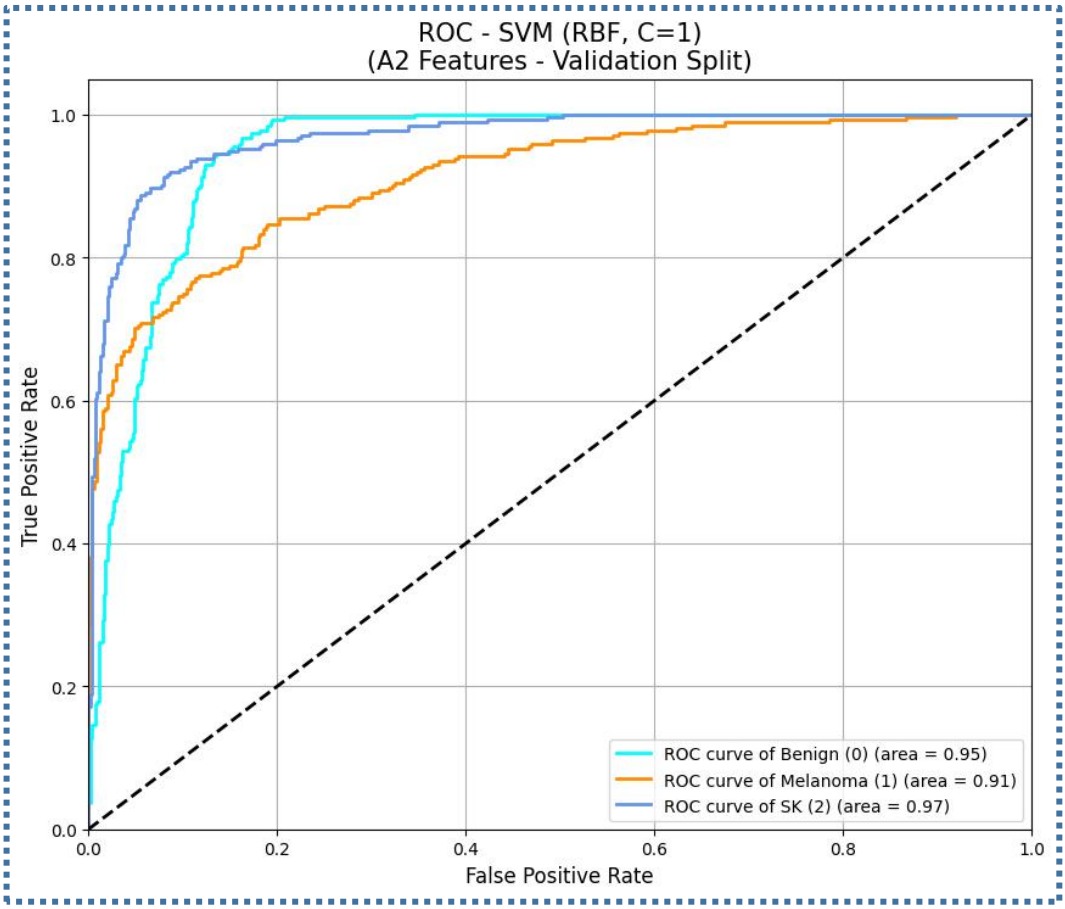
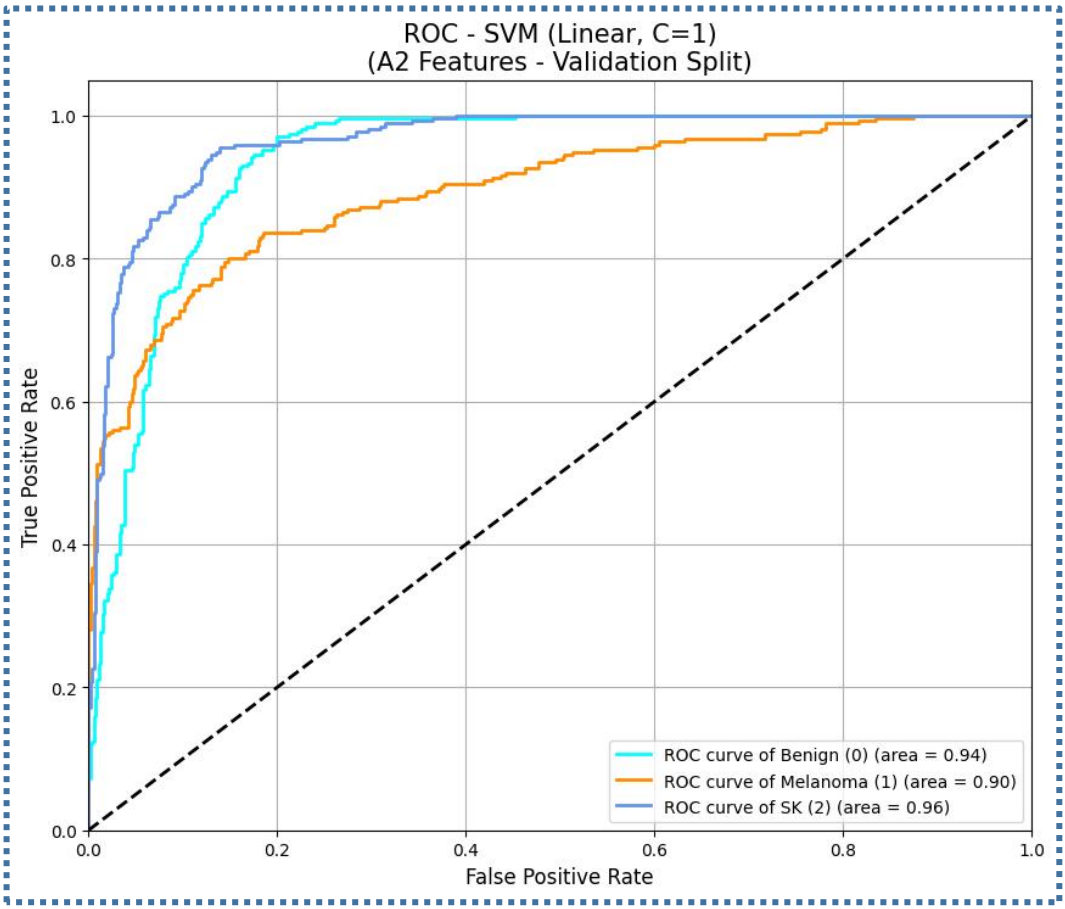
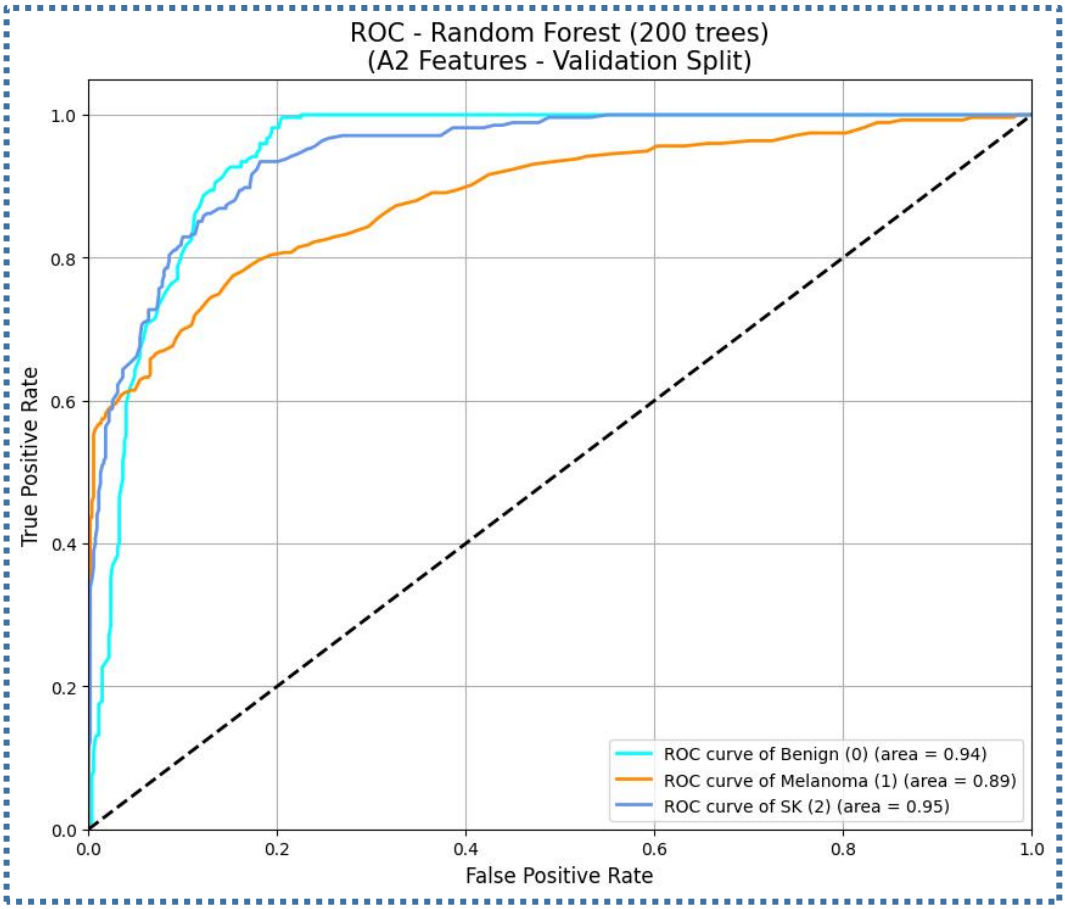
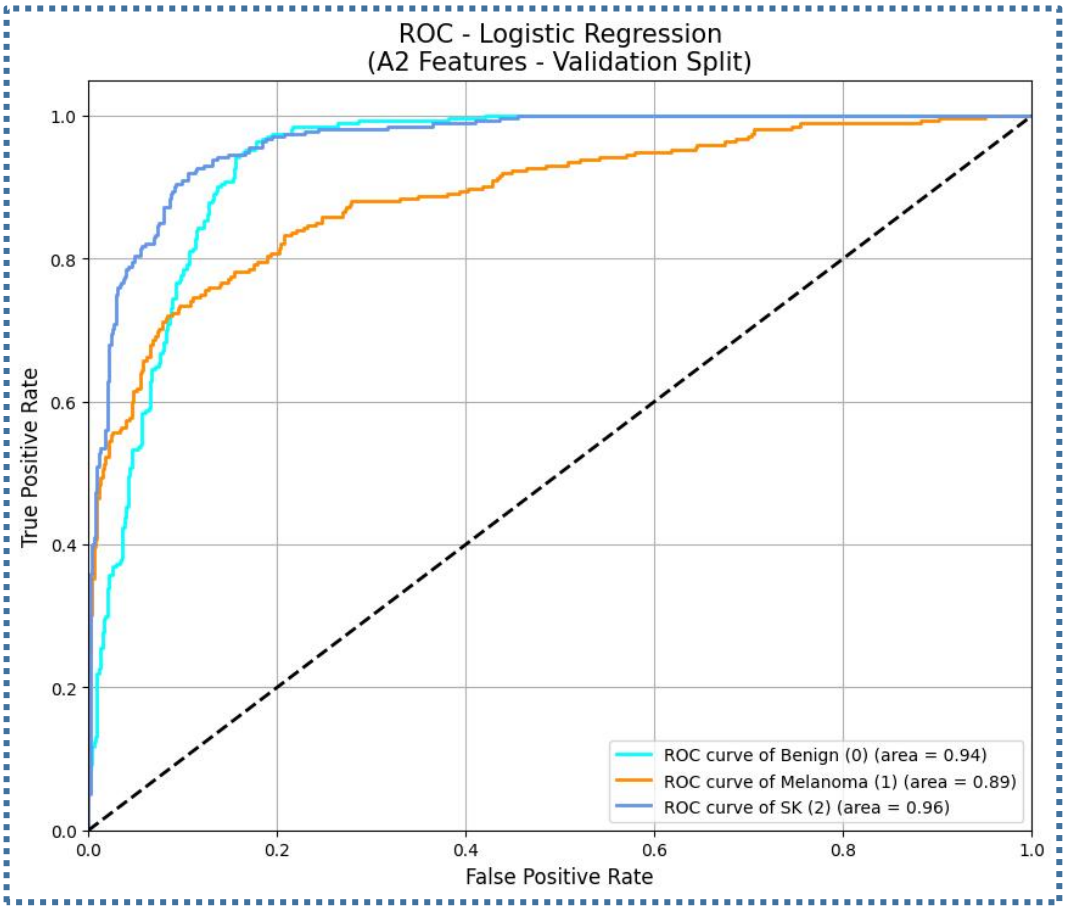


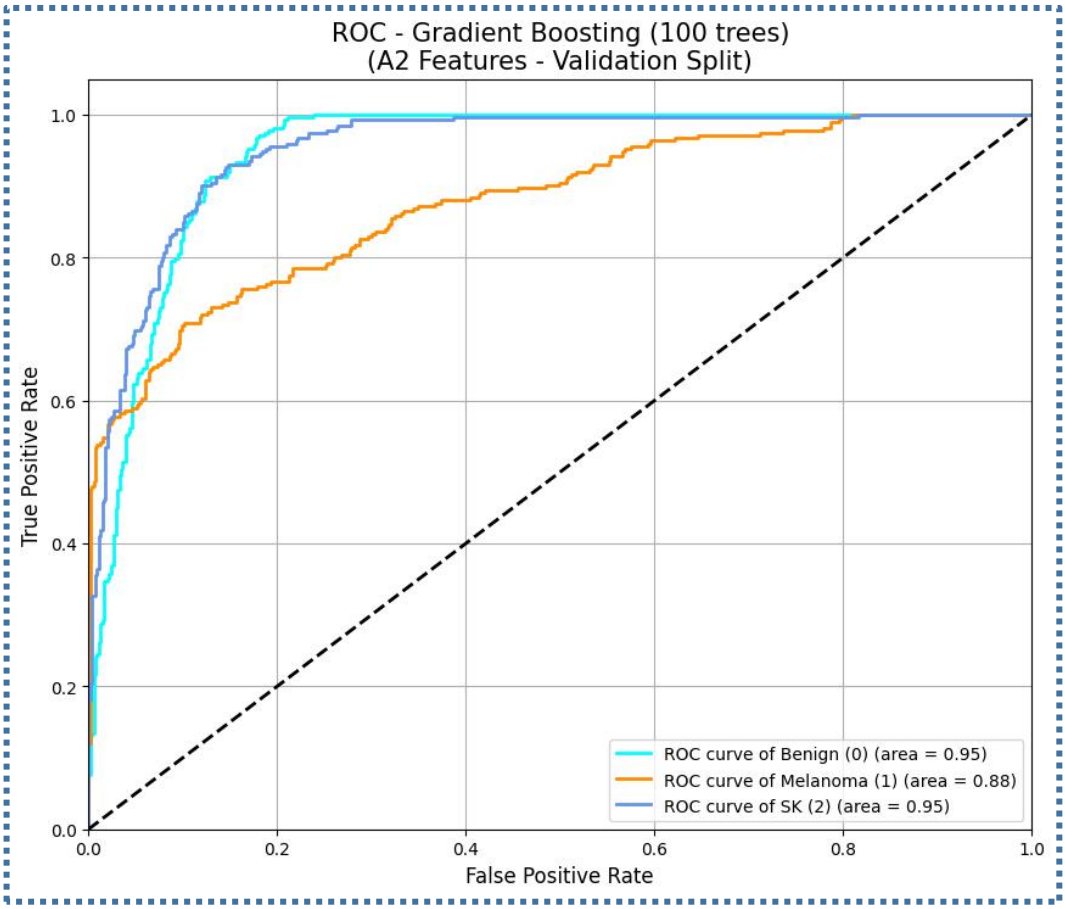
Figure 14 : Confusion Matrices for ML Classifiers with Combined LBP + EfficientNetB7 Features (Validation Split)

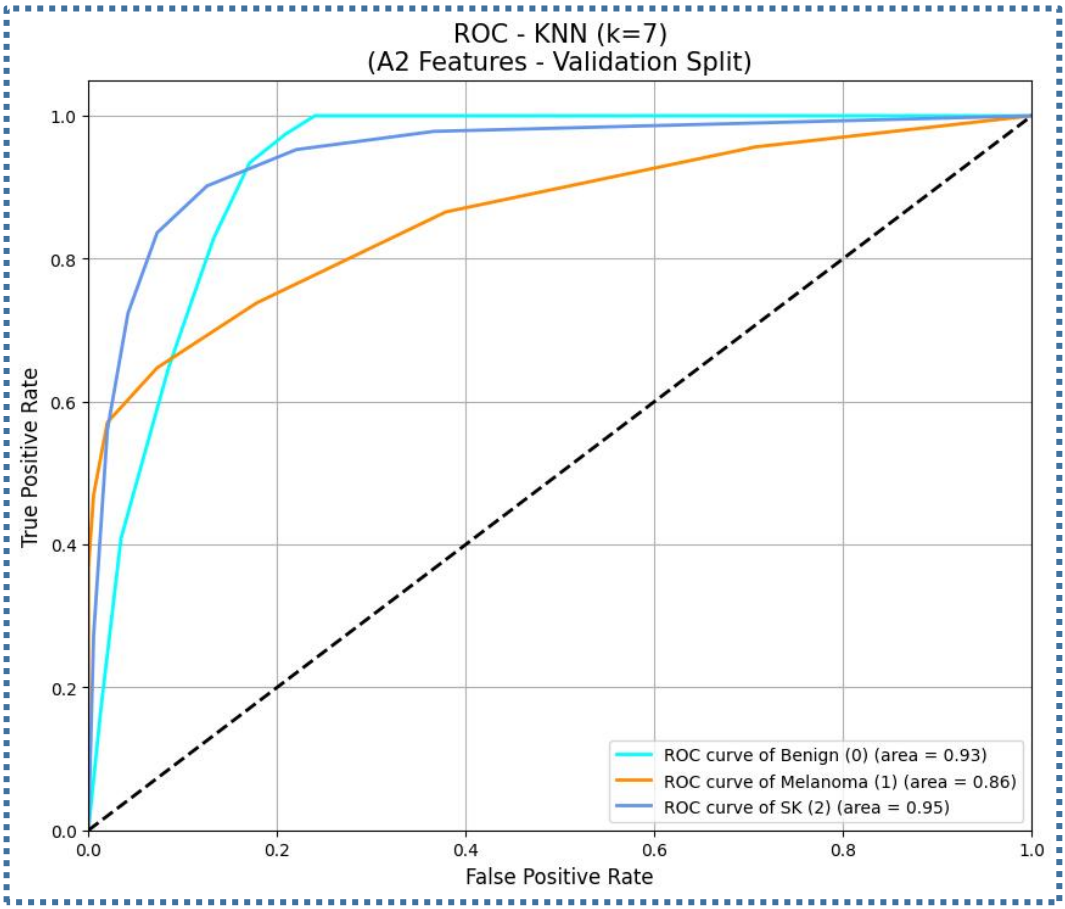

ROC - SVM (RBF, C=1)
(A2 Features - Validation Split)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of Benign (0) (area = 0.95)
ROC curve of Melanoma (1) (area = 0.91)
ROC curve of SK (2) (area = 0.97)


ROC - SVM (Linear, C=1)
(A2 Features - Validation Split)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of Benign (0) (area = 0.94)
ROC curve of Melanoma (1) (area = 0.90)
ROC curve of SK (2) (area = 0.96)


ROC - Random Forest (200 trees)
(A2 Features - Validation Split)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of Benign (0) (area = 0.94)
ROC curve of Melanoma (1) (area = 0.89)
ROC curve of SK (2) (area = 0.95)


ROC - Logistic Regression
(A2 Features - Validation Split)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of Benign (0) (area = 0.94)
ROC curve of Melanoma (1) (area = 0.89)
ROC curve of SK (2) (area = 0.96)


ROC - Gradient Boosting (100 trees)
(A2 Features - Validation Split)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of Benign (0) (area = 0.95)
ROC curve of Melanoma (1) (area = 0.88)
ROC curve of SK (2) (area = 0.95)


ROC - KNN (k=7)
(A2 Features - Validation Split)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of Benign (0) (area = 0.93)
ROC curve of Melanoma (1) (area = 0.86)
ROC curve of SK (2) (area = 0.95)

Hyperparameter Tuning: Similar GridSearchCV (3-fold CV) was used for SVM RBF, Logistic Regression, and Random Forest on the full 4116 scaled combined training features.

Evaluation of Tuned Models on Actual ISIC 2017 Test Set (600 images): Test set images were processed to extract combined LBP+EfficientNetB7 features and scaled.

Table 8 : Performance of Tuned ML Classifiers with Combined LBP + EfficientNetB7 Features (Actual Test Set)

Tuned Classifier	Accuracy	Benign (P/R/F1)	Melanoma (P/R/F1)	SK (P/R/F1)
Random Forest	0.6583	0.70/0.90/0.79	0.33/0.09/0.15	0.48/0.32/0.38
SVM (RBF)	0.6400	0.70/0.85/0.77	0.34/0.20/0.25	0.51/0.31/0.39
Logistic Regression	0.5967	0.72/0.72/0.72	0.28/0.29/0.28	0.47/0.44/0.45

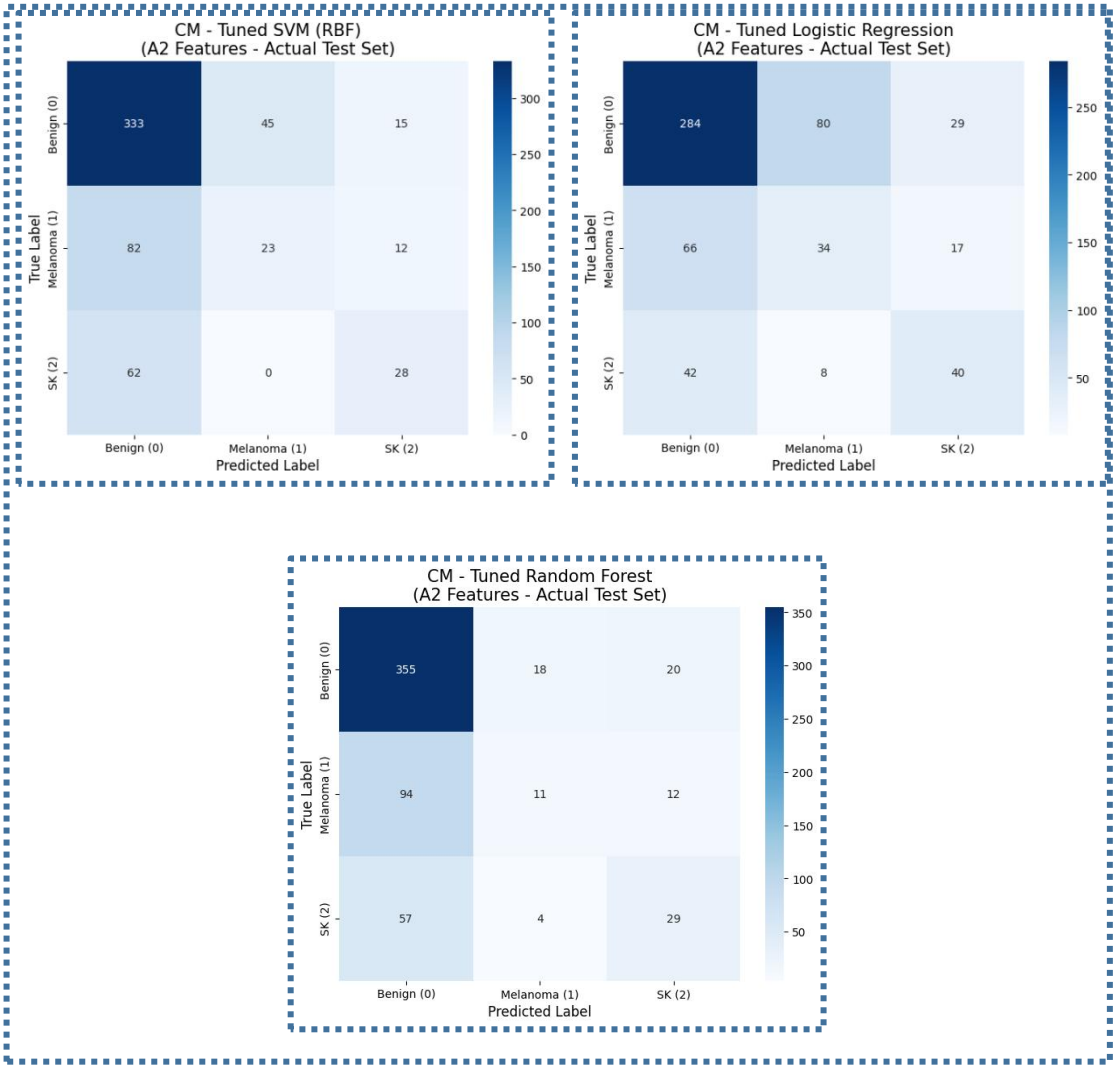
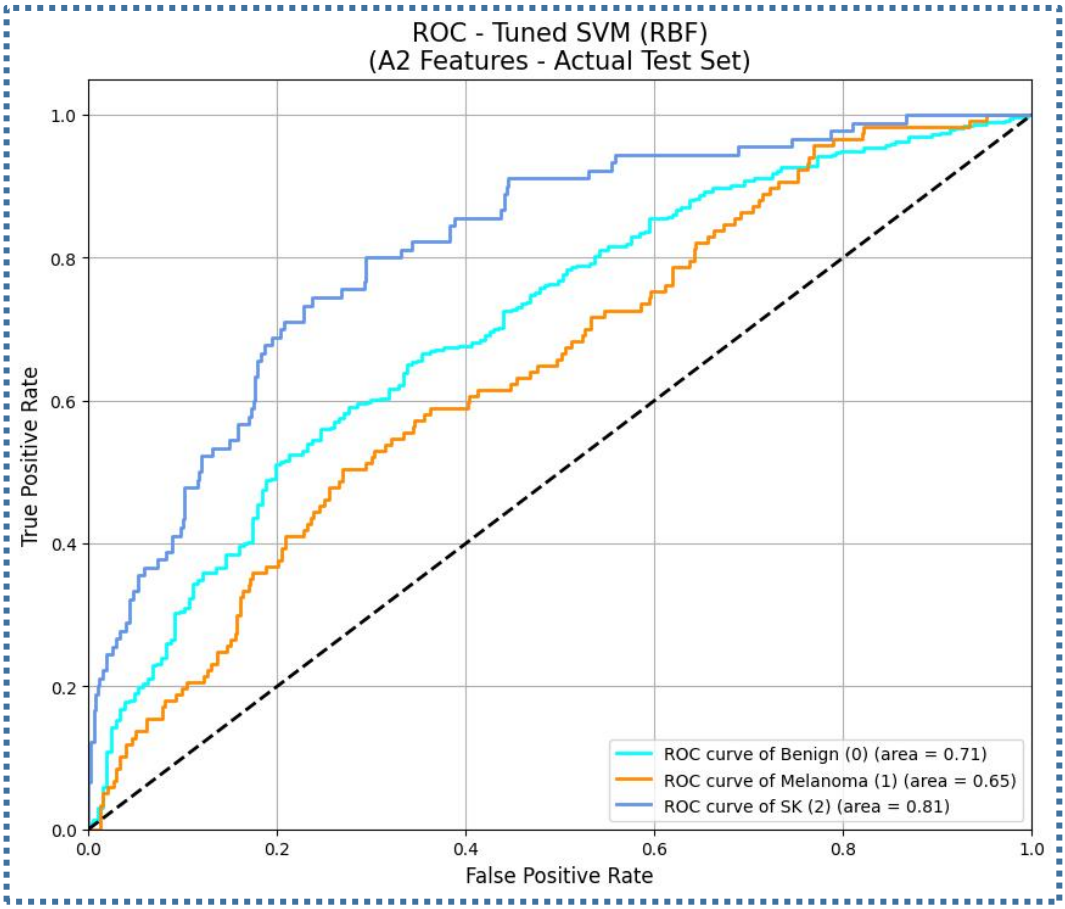
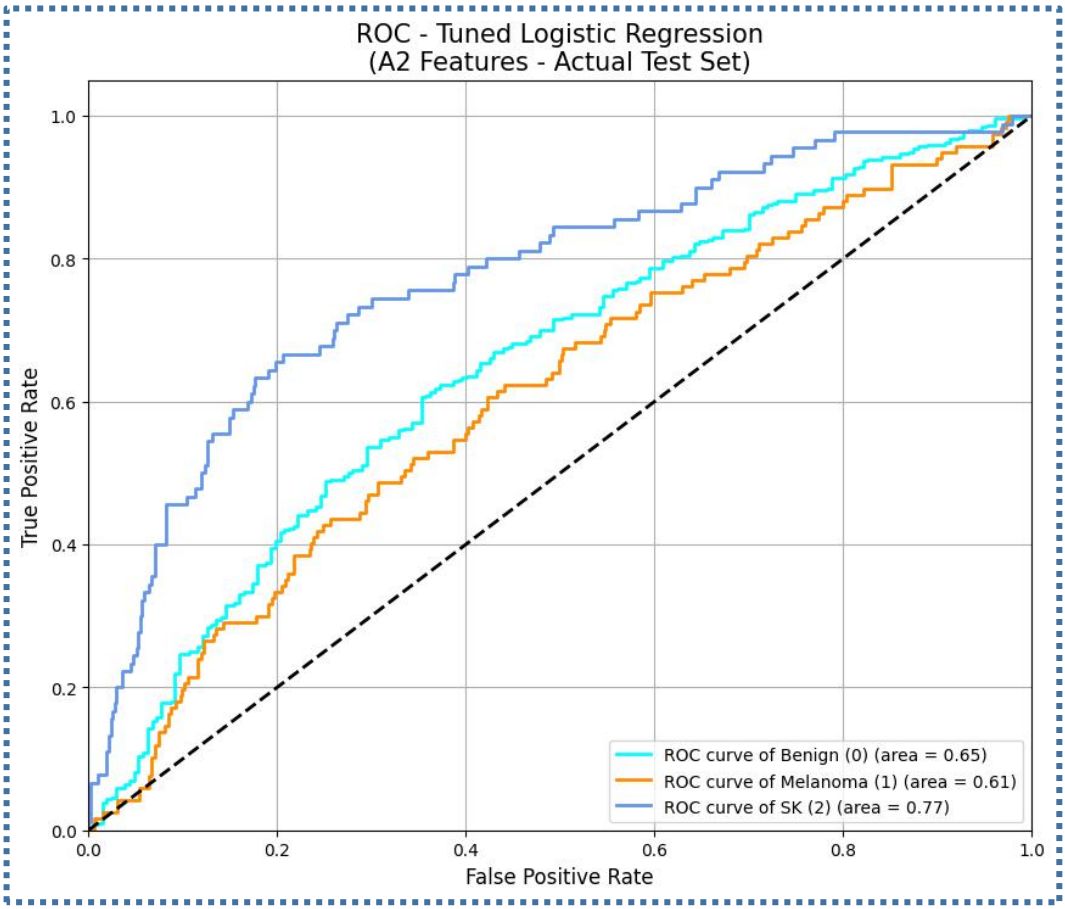
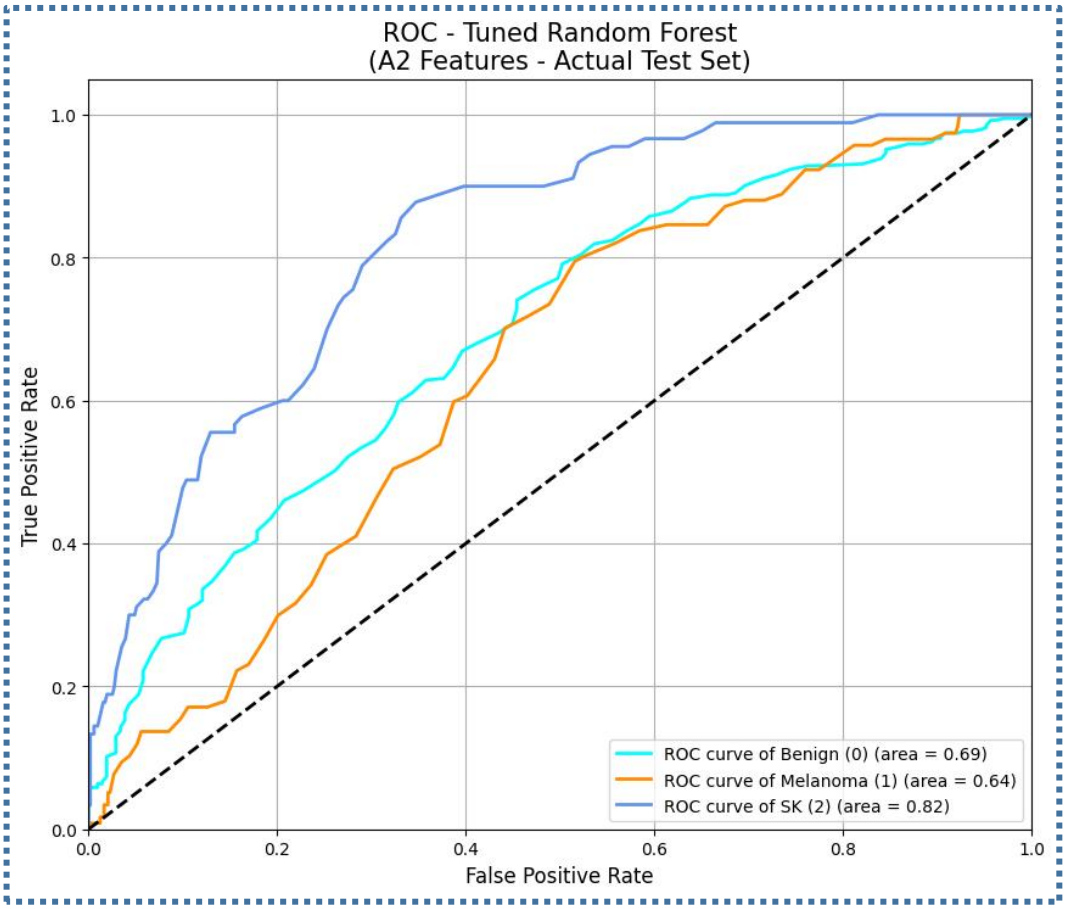


Figure 15 : Confusion Matrices for ML Classifiers with Combined LBP + EfficientNetB7 Features (Actual Test Set)


ROC - Tuned SVM (RBF)
(A2 Features - Actual Test Set)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of Benign (0) (area = 0.71)
ROC curve of Melanoma (1) (area = 0.65)
ROC curve of SK (2) (area = 0.81)


ROC - Tuned Logistic Regression
(A2 Features - Actual Test Set)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of Benign (0) (area = 0.65)
ROC curve of Melanoma (1) (area = 0.61)
ROC curve of SK (2) (area = 0.77)



4.5 Discussion of Results

The experimental results presented above provide valuable insights into the performance of various automated skin lesion segmentation and classification techniques. This section discusses these findings, compares different approaches, analyzes performance discrepancies, and acknowledges the challenges encountered.

4.5.1 Comparison of Segmentation Methods The segmentation experiments clearly demonstrated the superiority of deep learning-based methods over traditional approaches for this task.

- **U-Net** (Dice: 0.7697) and **SAM** (Dice: 0.7887, with manual prompts) significantly outperformed **Default MSER** (Dice: 0.0081).

- The **MSER+U-Net** hybrid (Dice: 0.6133) did not yield improvements over the standalone U-Net, likely due to the introduction of noisy or uninformative

MSER masks or a suboptimal feature fusion strategy.

- SAM's strong zero-shot performance is noteworthy, indicating the potential of foundation models, though practical application would require robust prompting strategies or fine-tuning.

4.5.2 Comparison of Feature Extraction Approaches for Classification:

- **Initial Hybrid (SAM-Segmented Features) vs. Revised Whole-Image Features:** The initial hybrid approach, using features from SAM-segmented regions, showed promise on the validation split (~81.19% accuracy) but failed drastically on the test set (~66.33% accuracy with 0% Melanoma recall). This highlighted issues with the stability/quality of features from variably segmented regions or the propagation of errors from the segmentation stage, especially when dealing with domain shift to the test set. Using whole-image features with EfficientNetB7 (Approaches 1 & 2) proved to be a more robust strategy for feature extraction, leading to better baseline performance on the validation data (up to ~84.10% accuracy).

- **EfficientNetB7 Only (Approach 1) vs. Combined LBP+EfficientNetB7 (Approach 2):** On the validation split, using EfficientNetB7 features alone (Approach 1) yielded slightly better results (best accuracy 84.10% with SVM RBF) than combining them with whole-image LBP features (Approach 2, best accuracy 82.52% with SVM RBF). This suggests that for data similar to the balanced training distribution, the deep features from EfficientNetB7 were sufficiently discriminative, and the addition of global LBP did not offer significant complementary information. On the actual test set, both approaches yielded very similar top accuracies (~65.7% - ~65.8%) with tuned Random Forest classifiers, with Approach 2 (combined features) showing marginally better recall for SK with some classifiers, though Melanoma recall remained critically low for both.

4.5.3 Performance of Different ML Classifiers:

- **On Validation Splits (Approaches 1 & 2):** SVM with an RBF kernel generally demonstrated the best performance among the classifiers evaluated with default parameters, achieving accuracies of 84.10% (Approach 1) and 82.52% (Approach 2). Logistic Regression also performed competitively.

- **On the Actual Test Set (Approaches 1 & 2):** After hyperparameter tuning, Random Forest classifiers emerged as slightly better performers in terms of overall accuracy (0.6567 for Approach 1, 0.6583 for Approach 2). However, this overall accuracy was significantly lower than on the validation set, and no classifier was able to effectively address the poor recall for Melanoma.

4.5.4 Impact of Hyperparameter Tuning (GridSearchCV & GA): Hyperparameter tuning was performed to optimize ML models for the (balanced) training data distribution. While this process likely improved performance on data drawn from that specific distribution (as suggested by GA validation scores around 84-85% for SVMs mentioned in the comprehensive report), it did not overcome the fundamental generalization gap when applied to the distinct and imbalanced ISIC 2017 Test Set. This suggests that the primary bottleneck was not suboptimal default ML parameters but rather the feature representation's robustness to domain shift and class imbalance.

4.5.5 Detailed Analysis of Performance Discrepancy (Validation vs. Test Set):

The most critical observation across all classification experiments was the substantial drop in performance when moving from the validation split (derived from the balanced, augmented training set) to the actual ISIC 2017 Test Set. This discrepancy is attributed to several interconnected factors:

1. **Domain Shift / Data Distribution Mismatch:** This is considered the most significant factor. The ISIC 2017 Test Set images, although from the same challenge, may possess subtle differences in acquisition conditions (e.g., cameras, lighting), patient demographics, or lesion characteristics compared

to the training set. Models trained meticulously on one distribution may not generalize well to another, even if related.

2. **Class Imbalance of the Test Set:** The training dataset was artificially balanced (1372 samples per class) via augmentation to prevent initial bias. However, the ISIC 2017 Test Set reflects a natural, imbalanced distribution (Benign: 393, Melanoma: 117, SK: 90). Models trained on perfectly balanced data often perform poorly when encountering such skewed distributions, as their decision boundaries are not optimized for it. This heavily explains the extremely low recall observed for Melanoma (e.g., 0.03-0.09% for some of the best test models) and SK, as the models became biased towards predicting the majority Benign class.
3. **Overfitting to the Training Data Distribution (including Augmentation Artifacts):** While validation sets help monitor overfitting during training, they are still derived from the same initial pool of images and thus share a similar underlying distribution, especially after augmentation has been applied to generate that balanced set. The models might have learned patterns or artifacts introduced by the augmentation process (e.g., specific orientations from rotations, edge effects from shifts/zooms) that were highly specific to this augmented training distribution but not representative of the un-augmented, real-world characteristics of the test set images.
4. **Image Size Consistency for Feature Extraction:** For Approaches 1 and 2, training images (already preprocessed and mostly 224x224 before augmentation pipeline) were effectively handled before being resized to 600x600 for EfficientNetB7 feature extraction. Test images, likely larger and more varied in original size, were directly downscaled to 600x600. This difference in handling and the more significant downscaling for test images might have introduced subtle variations in feature representation compared to the training set features.

4.5.6 Computational Challenges: The attempt to implement and fine-tune the Attention-Enhanced EfficientNetB7 model highlighted significant computational hurdles.

- Phase 2 of the fine-tuning (unfreezing backbone layers) with 600x600 input images proved infeasible on Google Colab's T4 GPU environment due to system RAM exhaustion, even with attempts to use small batch sizes.

- Difficulties were also encountered with local GPU (NVIDIA GeForce GTX 1050) detection or utilization for such large-scale tasks on the described PC. These challenges underscore the substantial computational resources required for training and fine-tuning state-of-the-art deep learning architectures on high-resolution images, which can be a limiting factor in research settings with constrained resources.

4.5.7 Limitations of the Study: The primary limitation of this study is the generalization gap observed between the validation performance on the balanced augmented training data distribution and the performance on the unseen, imbalanced test set. While various techniques were employed, robustly overcoming this domain shift and imbalance proved challenging with the explored feature extraction and classification pipelines. Furthermore, computational resource constraints prevented the full exploration of the planned two-phase fine-tuning for the attention-enhanced EfficientNetB7 model.

4.6 Chapter Conclusion

This chapter presented a detailed account of the experimental results from the various segmentation and classification methodologies investigated. In segmentation, deep learning models, particularly a trained **U-Net** (Dice: 0.7697) and the foundation model **SAM** (Dice: 0.7887 with manual prompts), demonstrated significantly superior performance compared to the traditional MSER algorithm on the ISIC 2017 Test Set.

For classification, approaches utilizing **EfficientNetB7 as a feature extractor** showed strong performance on validation splits derived from the balanced augmented training data, with an SVM (RBF kernel) achieving up to **84.10% accuracy** (Approach 1: EfficientNetB7 features only). Adding LBP features (Approach 2) did not yield further significant improvements on this validation data.

However, a critical finding was the substantial drop in performance across all classification models when evaluated on the independent, imbalanced **ISIC 2017 Test Set**. The best achieved accuracy on the test set was approximately **65.83%** (by a tuned Random Forest with combined LBP+EfficientNetB7 features, and similarly for EfficientNetB7 features alone with RF). More concerning was the extremely low recall for the clinically crucial Melanoma and Seborrheic Keratosis classes on this test set, indicating a strong model bias towards the majority Benign class.

The discussion highlighted that this performance discrepancy is primarily attributable to **domain shift** between the training and test distributions and the **impact of test set class imbalance** on models trained in a balanced environment, compounded by potential **overfitting to augmentation artifacts**. Computational limitations also constrained the full exploration of advanced fine-tuning strategies.

These findings underscore the significant challenges in developing CAD systems for skin lesion analysis that are not only accurate on curated training distributions but also robust and generalizable to real-world, imbalanced clinical data. The insights gained from these experiments pave the way for the overall conclusions of this thesis and recommendations for future research directions.

General Conclusion

This thesis embarked on the critical challenge of investigating and developing automated methods for skin lesion segmentation and classification from dermoscopic images, utilizing the publicly available ISIC 2017 dataset. The overarching goal was to contribute to the efforts aimed at improving early skin cancer detection, a crucial factor in enhancing patient outcomes. The research systematically explored a multi-stage pipeline, encompassing rigorous data preprocessing, extensive data augmentation to address inherent class imbalances, a comparative evaluation of various segmentation techniques, and an in-depth investigation into multiple feature extraction and classification strategies, including traditional machine learning approaches and advanced deep learning models.

The initial phase focused on meticulous **data preparation**. A comprehensive preprocessing pipeline involving hair removal, denoising, image enhancement, and resizing was implemented to standardize the ISIC 2017 training images. Recognizing the significant class imbalance within the dataset, an extensive data augmentation strategy was employed, successfully creating a balanced training set of 4116 images (1372 samples per class: Benign, Melanoma, and Seborrheic Keratosis). This balanced dataset formed the basis for training the subsequent classification models.

In the **segmentation experiments**, several methods were comparatively evaluated. Traditional MSER with default parameters proved inadequate for this complex task. Deep learning models, specifically a U-Net architecture trained on the dataset, demonstrated robust performance, achieving a Dice score of 0.7697. The foundational Segment Anything Model (SAM), even in a zero-shot setting with manual prompts, showed remarkable efficacy, yielding a Dice score of 0.7887. A hybrid MSER+U-Net model, however, underperformed compared to the standalone U-Net, likely due to the MSER input potentially introducing noise. These findings underscored the superiority of deep learning-based approaches for accurate lesion delineation.

The core of the thesis lay in the **classification experiments**. An initial hybrid approach, combining SAM-based segmentation with LBP and EfficientNetB7 features for SVM classification, showed promising results on a validation split derived from the balanced training data (accuracy ~81.19%). However, this model exhibited a severe performance degradation on the actual, imbalanced ISIC 2017 Test Set, with an accuracy of ~66.33% and a critical failure in recalling Melanoma cases. This prompted a shift towards using whole-image features. Two revised approaches were investigated: one using EfficientNetB7 features exclusively and another combining EfficientNetB7 features with LBP texture features. Both approaches yielded strong validation accuracies (up to ~84.10% with SVM RBF using EfficientNetB7 features only). Despite hyperparameter tuning, when evaluated on the ISIC 2017 Test Set, these models also experienced a significant performance drop, achieving a best accuracy of approximately 65.83% (with a tuned Random Forest using combined LBP and EfficientNetB7 features). Critically, the recall for minority classes (Melanoma and Seborrheic Keratosis) remained very low on the test set across all models. An attempt to fine-tune an Attention-Enhanced EfficientNetB7 model showed good initial validation performance (~84.59% in Phase 1) but was ultimately hindered by computational resource limitations, preventing full exploration.

A central finding of this thesis is the **significant challenge of model generalization** from a controlled, balanced, and augmented training environment to an unseen, imbalanced, real-world test set. The observed performance discrepancy highlights the critical impact of domain shift and the test set's natural class imbalance, which the models struggled to overcome despite achieving high validation scores.

The **contributions** of this thesis include: (i) the development and evaluation of a comprehensive preprocessing and augmentation pipeline tailored for the ISIC 2017 dataset; (ii) a comparative analysis of traditional, deep learning-based, and foundational models for skin lesion segmentation; (iii) an in-depth investigation of various feature extraction techniques (segmented vs. whole-image, LBP vs. deep features) and machine learning classifiers for lesion classification; (iv) an initial exploration of attention mechanisms in CNNs for this task; and (v) a crucial

empirical demonstration and analysis of the difficulties associated with domain generalization and class imbalance in the context of skin cancer classification.

This study faced certain **limitations**. The primary limitation remains the generalization gap of the classification models to the imbalanced test set, indicating that the features learned or the decision boundaries formed were not sufficiently robust to real-world data variations. Computational constraints also restricted the full fine-tuning of more complex models like the attention-enhanced EfficientNetB7, which might have offered improved performance. Furthermore, the SAM model for segmentation relied on manual prompts in this work, which would require an automated prompting strategy for practical deployment.

Building upon the findings and limitations of this research, several avenues for **future work** are proposed:

- **Advanced Domain Adaptation/Generalization Techniques:** Exploring methods specifically designed to improve model robustness to shifts between training and test data distributions.

- **Sophisticated Imbalance Handling:** Implementing more advanced techniques to address class imbalance during model training, such as specialized loss functions (e.g., focal loss, class-balanced loss), cost-sensitive learning, or more nuanced sampling strategies beyond simple oversampling.

- **Ensemble Methods:** Investigating the potential of combining predictions from multiple diverse models to improve overall accuracy and robustness.

- **Computationally Feasible Fine-Tuning:** Exploring strategies like knowledge distillation, parameter-efficient fine-tuning techniques for large models, or utilizing more optimized or lighter state-of-the-art deep learning architectures.

- **Automated Prompting for SAM:** Developing or integrating methods for automatic prompt generation to leverage SAM's powerful segmentation

capabilities in a fully automated pipeline.

- **External Validation:** Evaluating the best-performing models on diverse external datasets from different sources to more rigorously assess their generalization capabilities.

In conclusion, this thesis provided a thorough exploration of a multi-stage pipeline for automated skin lesion analysis using the ISIC 2017 dataset. While strong performance was achieved on validation data derived from a balanced training distribution, the work critically highlighted the substantial challenges of domain shift and class imbalance when applying these models to unseen, real-world test data. The insights gained underscore the complexity of developing truly robust and generalizable CAD systems for skin cancer classification and emphasize the need for continued research into methods that can effectively bridge the gap between controlled experimental setups and practical clinical application. This research contributes valuable empirical evidence and insights that can inform future efforts in this important domain of medical image analysis.

Bibiliography

[1] H. Sung, et al., "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians, 2024*.
Link:https://pubmed.ncbi.nlm.nih.gov/38572751/

[2] American Cancer Society, "Survival Rates for Melanoma Skin Cancer".
Link:https://www.cancer.org/cancer/types/melanoma-skin-cancer/detection-diagnosis-staging/survival-rates-for-melanoma-skin-cancer-by-stage.html

[3] A. Esteva, B. Kuprel, R. A. Novoa, et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, 2017.
Link:https://www.nature.com/articles/nature21056

[4] Cleveland Clinic, "Skin: Layers, Structure and Function."
Link: https://my.clevelandclinic.org/health/body/10978-skin

[5] DermNet, "Dermoscopy of melanoma."
Link:https://dermnetnz.org/cme/dermoscopy-course/dermoscopy-of-melanoma#:~:text=Dermoscopy%20of%20superficial%20melanoma,irregular%20in%20shape%20and%20structure.

[6] DermNet, "Dermoscopy of seborrhoeic keratosis."
Link:https://dermnetnz.org/cme/dermoscopy-course/dermoscopy-of-seborrhoeic-keratosis

[7] DermNet, "Dermoscopy of benign melanocytic lesions."
Link:https://dermnetnz.org/cme/dermoscopy-course/dermoscopy-of-benign-melanocytic-lesions

[8] S. W. Menzies, "Dermoscopy of pigmented skin lesions," *Postgraduate Medical Journal*, 2002.

Link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1742491/

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv*, 2014
Link:https://arxiv.org/abs/1409.1556

[10] K. He, et al., "Deep residual learning for image recognition," in *Proc. CVPR*, 2016
Link:https://arxiv.org/abs/1512.03385

[11] N. C. F. Codella, et al., "Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI)...," *arXiv*, 2018.
Link:https://arxiv.org/abs/1710.05006

[12] A. Menegola, et al., "RECOD Titans at ISIC Challenge 2017," *arXiv*, 2017.
Link: https://arxiv.org/abs/1703.04819

[13] A. Nayak, et al., "A patch-based deep learning architecture for the classification of skin lesions," *Biomedical Signal Processing and Control*, 2024.
Link:https://www.sciencedirect.com/science/article/pii/S174680942300582X

[14]:P. Prakash, et al., "MFFDCNN-CTDC: Multi-Scale Feature Fusion of Deep CNNs for Cancerous Tumor Detection and Classification in Histopathological Images," *IEEE Journal of Biomedical and Health Informatics*, 2024.
Link: https://ieeexplore.ieee.org/abstract/document/10424564

[15]:O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv*, 2015.

Link:https://arxiv.org/abs/1505.04597

[16]: A. Kirillov, et al., "Segment Anything," *arXiv*, 2023.
Link: https://arxiv.org/abs/2304.02643

[17] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *arXiv*, 2019.
Link: https://arxiv.org/abs/1905.11946