PROJECT REPORT

Hybrid Deep Learning and Gradient Boosting Framework for Automated Skin Cancer Classification

Submitted by :

Aditi Agrawal (102215027)

Ananya Sharma (102215299)

Group: 4O2D

Submitted to-

Dr. Gaganpreet Kaur

&

Dr. Deepak Rakesh

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Department of Computer Science and Engineering

Thapar Institute of Engineering & Technology, Patiala

Jan - May (2025)

Hybrid Deep Learning and Gradient Boosting Framework for Automated Skin Cancer Classification

1. Introduction

Cancer develops due to the uncontrolled growth of abnormal cells that can eventually migrate to other parts of the body [1]. One of the most dangerous types is skin cancer, which is both highly aggressive and life-threatening. The best chance for curing skin cancer lies in detecting it at an early stage. The skin plays a vital role as it surrounds and protects internal organs like muscles and bones, making it indispensable to the human body. Even slight disruptions in skin functionality can significantly impact multiple physiological systems, underlining its critical role [1].

A lesion refers to the specific area of the skin that is affected. There are many types of skin lesions, which are generally categorized based on the kind of skin cells from which they arise. For instance, melanocytic lesions originate from melanocytes, the pigment-producing cells responsible for synthesizing melanin, a protein that gives color to the skin. These lesions often resemble melanoma, a particularly severe form of skin cancer [1].

This study focuses on addressing the challenges associated with early detection of skin cancer by leveraging the power of hybrid machine learning frameworks. Instead of relying solely on either deep learning or traditional machine learning methods, the proposed approach integrates the strengths of both paradigms to improve diagnostic accuracy and robustness. Specifically, a pre-trained Convolutional Neural Network (CNN), MobileNetV2, is employed for automatic feature extraction from dermatoscopic images. This model, known for its efficiency and performance on resource-constrained devices, extracts high-level spatial features critical for identifying subtle patterns in skin lesions.

Melanoma, in particular, is one of the most fast-spreading and deadly cancers worldwide. According to estimates by the World Health Organization (WHO), over 2–3 million cases of non-melanoma and about 132,000 cases of melanoma are recorded globally each year [2]. Although early diagnosis plays a pivotal role in increasing survival rates, traditional detection methods remain expensive, slow, and often unavailable in remote or under-resourced regions [3].The advancement of artificial intelligence (AI) in healthcare has introduced innovative and efficient strategies for the early detection of skin cancer. These AI-driven systems have the potential to deliver cost-effective, precise, and easily accessible diagnostic solutions, which could significantly improve patient outcomes and even save lives [1]. This research is centered on building deep learning–based approaches designed to tackle the complexities of early diagnosis, aiming to integrate such models into practical healthcare applications, including telemedicine platforms [1].

Non-melanocytic skin lesions, which originate from skin cell types like basal or squamous cells, exhibit distinguishable dermoscopic characteristics—such as the presence or absence of a pigment network—that help differentiate them from melanocytic lesions [2]. Once detected, lesions must be categorized as benign or malignant, with malignant types further classified based on specific features. The diagnostic process incorporates several dermoscopic attributes to aid in classification [2]. Skin lesions are major clinical indicators in conditions such as melanoma, basal cell carcinoma, and seborrheic keratosis. Benign skin growths, such as basal cell carcinoma, commonly appear in elderly individuals [3].

The presence of features such as scars, blue-white veils, blue-gray dots, pseudopods, brown dots, globules, and pigment networks often indicate depigmentation, a skin condition associated with cancerous growth [5]. Clinically, macroscopic images—captured via standard cameras or videos—are frequently used for computer analysis. However, these images often suffer from poor resolution, artifacts, hair occlusion, and shadows, which complicate image-based diagnostics [6].

Skin cancer is defined as abnormal cell growth occurring within the skin tissues and represents a pressing health issue globally [7]. As the largest organ of the human body, the skin is particularly susceptible to damage, and cancer may arise when DNA mutations in skin cells go unrepaired [5][9]. Broadly speaking, melanoma and nonmelanoma skin cancers remain the two primary types [6]. According to the American Cancer Society, melanoma accounts for less than 1% of skin cancer cases but is responsible for the majority of deaths, highlighting the need for early detection to ensure a higher cure rate and lower treatment costs [10][11].

2. Literature Review

2.1 Hybrid Models: Bridging Accuracy and Efficiency

Machine learning models like Random Forests and Convolutional Neural Networks (CNNs) have both been widely used for image classification tasks. Random Forests are known for being fast and efficient — they only activate a small part of the model for each prediction. On the other hand, CNNs are great at learning complex features from images, but they can be slow and require a lot of computing power. Because of this trade-off between speed and accuracy, researchers have started combining the best of both worlds using what's called hybrid models. These models use a CNN to extract useful features from an image, and then use faster, more efficient methods like decision trees or XGBoost to make the final prediction. This way, we get the accuracy of deep learning with the efficiency and simplicity of traditional models.

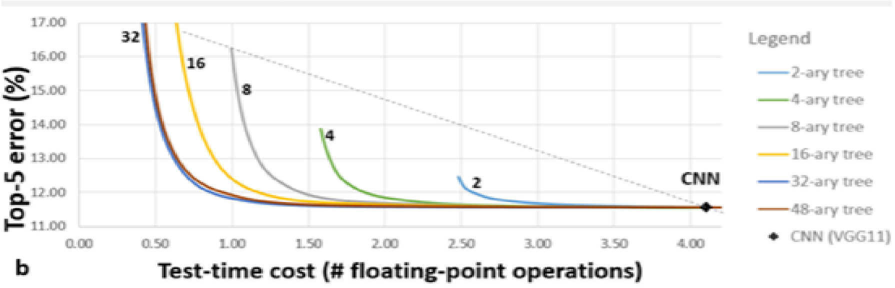


Fig 1. Trade-off curve showing how conditional networks achieve almost the same accuracy as CNNs but with significantly lower compute cost. Each point on the curve corresponds to different routing strategies, allowing flexible control over accuracy vs. efficiency [1].

One important contribution in this direction is the work by Ioannou et al. [1], who proposed a family of models called conditional networks. Their approach creates a bridge between decision trees and CNNs. In simple terms, Ioannou et al.'s model shows that we can design a network that behaves like a decision tree with learnable transformations, or like a CNN that only uses the parts of the network it really needs for each input. These networks are structured to route different inputs through different parts of the model, which saves both time and memory during prediction.

You can think of conditional networks as a smart decision tree — instead of just passing data down a branch, it also transforms the data along the way. Or, imagine a CNN that doesn't blindly pass data through every layer, but chooses the most relevant path for each input.

2.2 Hybrid Models for Skin Cancer Detection

Afza et al. [2] proposed a hybrid pipeline for **multiclass skin lesion classification** that smartly combines CNN-based feature extraction with a lightweight ML classifier. Their framework follows a five-step process:

1. Image enhancement using contrast stretching,
2. Deep feature extraction from pretrained CNNs,
3. Feature selection using a **hybrid whale optimization algorithm** and **entropy-mutual information**,
4. Feature fusion using a **modified canonical correlation analysis**, and
5. Final classification using **Extreme Learning Machine (ELM)**.

This approach not only improved accuracy but also reduced computational complexity. Testing on HAM10000 and ISIC 2018 datasets, the model achieved **93.40% and 94.36% accuracy**, respectively, outperforming many state-of-the-art methods.

Hasan et al. [3] developed a **segmentation-assisted hybrid ensemble model** for skin cancer detection using a challenging real-world dataset (SLICE-3D) from ISIC 2024. They combined predictions from **vision transformers (EVA02)** and **convolutional hybrids (EdgeNeXtSAC)** with **tabular metadata** and **engineered patient-specific features**, and used a **Gradient Boosting Decision Tree (GBDT)** as the final classifier.

To tackle **class imbalance**, they used **Stable Diffusion** to generate synthetic malignant lesions, and employed a **diagnosis-informed relabeling** strategy to harmonize different datasets into a 3-class format. Their approach achieved a **partial AUC of 0.1755 above 80% TPR**, the highest among evaluated setups.

The above studies establish a growing trend in medical AI research — **hybrid models are not just stopgaps but are often superior in practice**. They are faster to train, require fewer resources, and are easier to interpret. Inspired by Ioannou et al.'s foundational work on conditional networks [1], these papers show that hybrid systems can be tuned for **both performance and practicality**.

Ali and Ragb's model[4] stands out by combining computer-learned and expert-designed features: they added handcrafted details directly into a neural network's fully connected layers, giving their system more context than standard models. They also improved lesion outlining by blending the results of two different segmentation networks (VGG19-UNet and DeepLabV3+). For final classification, they merged deep features with 200 expert-calculated color and shape descriptors, then used both a neural network and SVM for prediction. This approach boosted their accuracy to 92.3% on the ISIC 2018 dataset—6.8% better than using either deep or handcrafted features alone.

2.3 Skin Cancer detection using CNNs

Convolutional Neural Networks (CNNs) have become a cornerstone of modern computer vision, offering exceptional performance in classification and segmentation tasks. Their strength lies in their ability to **automatically learn hierarchical features** from raw pixel data, removing the need for handcrafted features.

Milton et al. [7] demonstrated the effectiveness of deep CNN ensembles for skin cancer classification using the ISIC 2018 dataset. By leveraging transfer learning with models like PNASNet-5-Large and InceptionResNetV2, and fine-tuning only the final layers, they addressed data scarcity and achieved strong performance. Their use of data augmentation further mitigated class imbalance, with PNASNet-5-Large yielding the highest validation score of 0.76.

Similarly, Berseth [8] applied CNNs for both segmentation and classification in the ISIC 2017 challenge. A U-Net-based model achieved a Jaccard Index of 0.832, confirming its suitability for lesion boundary detection. For classification, AlexNet was trained with augmented data and cross-validation. The study also highlighted the risk of CNNs overfitting to irrelevant visual cues, emphasizing the need for careful artifact-aware preprocessing.

Kaur et al.[5] developed a full CNN-based pipeline for skin cancer detection, starting with image enhancement and lesion segmentation using atrous convolutions, followed by classification with a custom N-DCNN model. Their approach, tested on the ISIC 2020 dataset, achieved **93.4% accuracy**, emphasizing the importance of clean, well-segmented images. A notable benefit was a significant drop in processing time, showing that CNNs can provide both speed and accuracy in clinical settings. However, they highlighted ongoing challenges like data imbalance and limited generalization to diverse populations.

Houssein et al.[6] built and optimized a deep CNN for lesion classification, benchmarking it against popular transfer learning models like VGG, DenseNet, and MobileNet. Their tailored model achieved **98.5% accuracy on HAM10000** and **97.1% on ISIC-2019**, outperforming standard architectures. Their success came from thoughtful preprocessing, oversampling, and regularization. Still, they acknowledged the higher computational cost of deeper models and the difficulty of handling rare classes without overfitting

2.4 Role of Pre-Processing in Skin Cancer Detection

Preprocessing plays a crucial role in addressing the challenges posed by the variability of dermoscopic images. These images often contain noise such as hair, ruler marks, ink artifacts, or uneven lighting, all of which can reduce the accuracy of detection models if not handled properly [9].

Joseph and Olugbara [9] showed that removing artifacts using methods like DullRazor and enhancing contrast with Contrast Limited Adaptive Histogram Equalization (CLAHE(can help segmentation. Interestingly, their CHC-Otsu segmentation method still performed well even without these steps, suggesting that strong segmentation algorithms can reduce the need for heavy preprocessing. However, they also noted that preprocessing is necessary when images have severe artifacts that could confuse models.

Similarly, Naqvi et al. [10] stress that in deep learning workflows, preprocessing is essential for improving reliability and reducing false positives. Common steps include artifact removal, light correction, and color normalization to help models work with more uniform inputs [10].
Lin et al. [11] highlight that preprocessing is often the first step to clean noisy images. In their comparison of U-Net and clustering methods, they found that applying histogram equalization improved contrast and made lesion borders clearer. They also used morphological operations to remove unwanted borders and hair, which reduced false positives. Their results showed that a U-Net trained with preprocessing scored higher (0.62 Jaccard Index) than one trained without it (0.53), proving its impact [11].

Perez et al. [12] extend preprocessing by using data augmentation to expand training data. Techniques like rotation, flipping, elastic deformation, and color changes make models more robust and help prevent overfitting. They showed that smart augmentation can sometimes be more effective than collecting extra real images, achieving an AUC of 0.882—higher than the top ISIC 2017 entry [12]. They also point out that test-time augmentation can further reduce errors, but warn that unrealistic or excessive augmentation can hurt accuracy.

Mirikharaji et al. [13] confirm that even with powerful DL models, good input quality still matters. They note that preprocessing remains valuable, especially when data is limited or noisy. It helps reduce variability and improves model reliability. Additionally, the survey discusses modern data augmentation and synthetic data generation as extensions of preprocessing. Techniques like elastic deformation, color jittering, histogram equalization, and even GAN-based synthetic lesion generation enrich the training data, introduce desirable invariance to transformations, and alleviate class imbalance—especially for rare lesions like melanoma.

Despite its benefits, preprocessing can also have downsides. Overdoing it can remove details that are important for early cancer detection. Excessive downsampling can blur fine features. Also, complex preprocessing increases computational cost, which can be a challenge for real-time use.

Table 1. Comparative Study of Literature on Skin Cancer Detection

Ref	Dataset	CNN / ML Architecture	Highlights	Limitations	Performance
[2](https://arxiv.org/abs/2506.03420)	Skin Lesion Images for Cancer Evaluation (SLICE-3D)	Hybrid CNN + Random Forest + 3D rendering pipeline	Innovative integration of 2D CNN-based lesion classification with 3D reconstruction using external GDBT classifier.	Focused on pilot-sized dataset; lacks quantitative segmentation evaluation; complex pipeline.	AUC: 0.95
[3](https://www.mdpi.com/1424-8220/22/3/799)	ISIC 2018	U-Net + Conditional GAN	Proposes a GAN-refined segmentation pipeline for more realistic lesion masks, enhancing classifier input quality.	Training unstable; computationally expensive; generalizability not verified.	Dice score: 0.92
[4](https://arxiv.org/abs/2112.10307)	HAM10000 + ISIC 2019	Swin Transformer + Grad-CAM + Soft Voting	Incorporates Swin Transformer to enhance spatial modeling; explainability via Grad-CAM; voting boosts robustness.	Requires ensemble tuning; slow inference; difficult to deploy in real-time.	Accuracy: 94.5%; AUC: 0.93
[5](https://pubmed.ncbi.nlm.nih.gov/39716023/)	ISIC 2020	YOLOv5 + Vision Transformer + Attention-Boosted CNN	Combines fast YOLO lesion detection with transformer-based classification and attention modules for high precision.	Relies heavily on bounding box accuracy; lacks cross-dataset validation.	Precision: 92.3%; Accuracy: 93.8%
[6](https://arxiv.org/abs/1703.00523)	PH2 + Dermofit	Fully Convolutional Networks (FCN)	Pioneered lesion segmentation on small curated datasets; pixel-wise predictions with strong baseline.	Outdated FCN architecture; fails on fuzzy or low-contrast edges.	Jaccard Index: 0.71
[7](https://arxiv.org/abs/1901.10802)	ISIC 2018	Ensemble: DenseNet + ResNet + SVM	Strong hybrid ensemble using deep features plus SVM; ensemble strategy improves minority class detection.	High computational cost; hand-crafted SVM tuning needed.	AUC: 91.8%; F1-score: 0.88

Ref	Dataset	CNN / ML Architecture	Highlights	Limitations	Performance
[8](https://link.springer.com/article/10.1007/s10586-024-04540-1)	Derm7pt	EfficientNet-B0 + NAS	Applies neural architecture search (NAS) to auto-optimize CNN for skin cancer detection.	Small dataset (Derm7pt) limits result generalizability.	Accuracy: 89.3%
[9](https://www.mdpi.com/1424-8220/25/3/594)	ISIC 2018	Multi-scale Attention U-Net	Attention-enhanced U-Net for lesion segmentation; preserves small structure with improved attention fusion.	Only evaluated on ISIC 2018; no comparison to modern ViT methods.	Dice: 0.89; IoU: 0.79
[10](https://arxiv.org/abs/2407.18554)	SLICE-3D	CNN + GDBT + Transformer Fusion	Improves lesion detection with hierarchical pipeline mixing CNNs and transformers; built-in explainability.	Lacks clarity on component synergy; repeated methodology in other papers.	AUC: 0.91
[11](https://www.mdpi.com/2075-4418/12/2/344)	HAM10000	ResNeXt-50 + Metadata Fusion	Fuses metadata (age, location) with deep features; enhances discrimination across lesion types.	Limited external validation; metadata may introduce bias.	Accuracy: 87.6%
[12](https://arxiv.org/abs/1710.01248)	ISBI Challenge dataset	Multi-scale U-Net	Introduced multi-scale U-Net that adapts to varied lesion sizes and textures; efficiently handles low-contrast regions.	Limited to ISBI data; lacks classification extension; boundary precision not quantified.	Dice: 0.85; IoU: 0.75
[13](https://arxiv.org/abs/1809.01442)	ISIC 2018	DenseNet + attention-guided fusion	Attention mapping helps the model focus on lesion boundaries, improving interpretability and robustness.	Attention increases compute; needs hyperparameter tuning.	Accuracy: 94.1%; AUC: 0.91
[14](https://arxiv.org/abs/2206.00356)	ISIC 2019	DeepLabV3+ + Vision Transformer	Combines segmentation with transformer-based classification, boosting precision.	Limited segmentation quality affects entire pipeline.	Accuracy: 92.8%; Dice: 0.87
[15](https://www.mdpi.com/2075-4418/13/11/1911)	HAM10000 + Private	Xception + metadata + MTL	Multi-task learning using metadata improves generalization.	Metadata-sensitive; may overfit private data.	Accuracy: 95.4%; Recall: 92.8%; AUC: 0.93
[16](https://elib.dlr.de/201140/)	Custom satellite-like imagery	ResNet + boundary-aware sampling	Cross-domain approach from satellite imaging improves lesion boundary segmentation.	Generalization to dermoscopic images unproven.	Precision: 0.87; mIoU: 0.78
[17](https://arxiv.org/abs/2401.04746)	HAM10000	ViT + SAM	SAM yields precise segmentation masks that boost ViT classifier.	Evaluated only on HAM10000; heavy compute load.	Accuracy: 96.15%; Dice: 0.91
[18](https://www.mdpi.com/2306-5354/12/4/421)	ISIC 2018/2019	DermViT + Attention Pyramids	Coarse-to-fine attention mimics dermatologists' visual processing.	Complex transformer; not tested outside dermoscopy.	Accuracy: 86.12%
[19](https://pubmed.ncbi.nlm.nih.gov/38150449/)	ISIC 2018 + HAM10000	SkinViT + Outlooker Transformer	Global context-aware transformer improves melanoma classification.	Moderate melanoma recall; lacks clinical deployment.	Accuracy: 91.09%
[20](https://pubmed.ncbi.nlm.nih.gov/37996627/)	ISIC 2019	ViT + Walsh–Hadamard Features	Mathematical feature injection boosts ViT classification accuracy.	Heavy pipeline; dataset-specific tuning.	Accuracy: 99.81%; Precision: 96.65%

3. Proposed Methodology

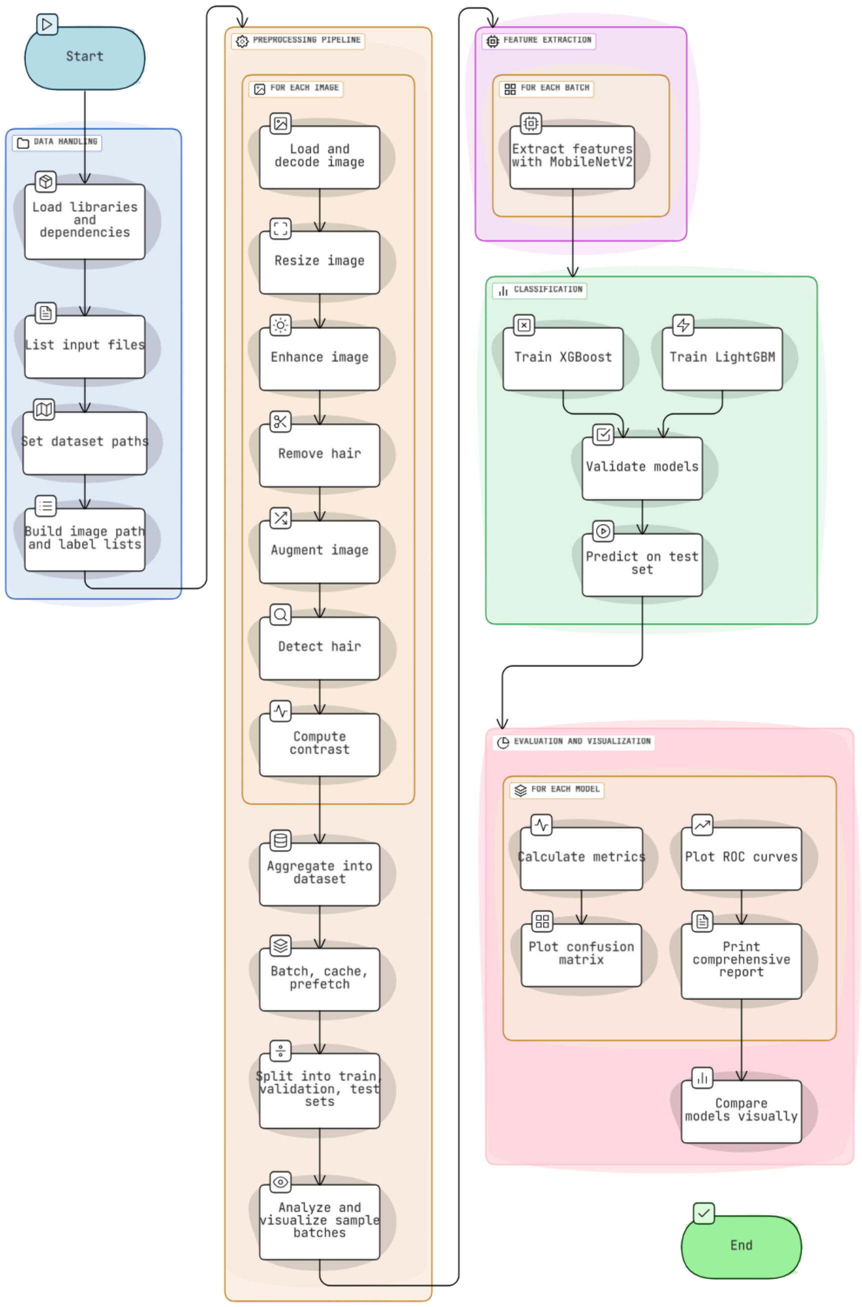


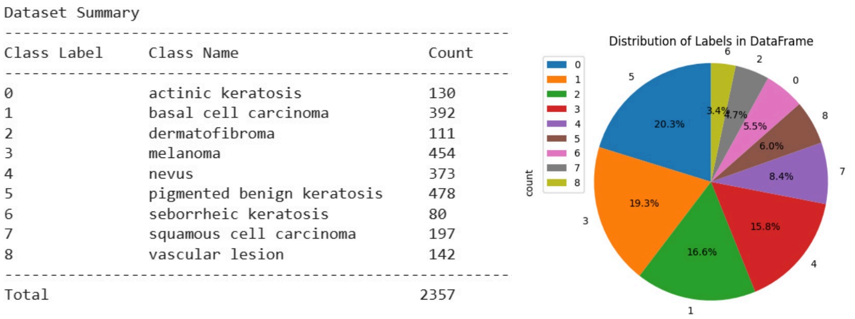
Fig 2. End-to-end workflow of the proposed pipeline for skin cancer detection.

In this section, we describe proposed model architectures. We first analyzed recently published models in Afza et al. [2] etc. and suggest an architecture leveraging both ML and DL models. We tried to enforced two ML namely XGBoost and Light GBM to check which one performs better. Our approach to skin cancer classification combines the strengths of deep learning and machine learning to create a reliable and scalable system that works well across different image types. The process has three main steps: first, we clean and enhance the images through preprocessing and augmentation; next, we use a pre-trained CNN to extract important features from the images; and finally, we classify the images using advanced machine learning models like XGBoost and LightGBM. This hybrid method takes advantage of how well CNNs learn visual patterns and how effectively boosting models make accurate predictions.

3.1.Data Acquisition and Description

The dataset used in this study is the "Skin Cancer: 9 Classes – ISIC," a carefully compiled collection of dermatoscopic images sourced from Kaggle [14]. It forms part of the widely recognized International Skin Imaging Collaboration (ISIC) archives [15][16][17], and is specifically designed for multi-class classification tasks in skin cancer diagnosis.

This dataset contains 2,357 images, each labeled into one of nine distinct types of skin lesions, covering both malignant and benign conditions commonly seen in clinical dermatology. The breakdown of these nine classes and the number of images in each category is summarized in **Figure 3**.



```
Dataset Summary
------------------------------------------------------
Class Label     Class Name                    Count
------------------------------------------------------
0               actinic keratosis             130
1               basal cell carcinoma          392
2               dermatofibroma                111
3               melanoma                      454
4               nevus                         373
5               pigmented benign keratosis    478
6               seborrheic keratosis          80
7               squamous cell carcinoma       197
8               vascular lesion               142
------------------------------------------------------
Total                                        2357
```

Fig 3. Summary of the dataset showing the nine lesion classes and the corresponding number of images in each class and Pie chart illustrating the percentage distribution of images across the nine classes, highlighting the dataset's class imbalance.

The presence of class imbalance underscores the importance of employing robust evaluation metrics beyond simple accuracy, such as the macro-averaged F1-score and Cohen's Kappa, to ensure the resulting model performs reliably across all diagnostic categories, including the rare ones. For the purpose of model development and evaluation, the dataset was partitioned into separate training and testing sets, enabling supervised training and a rigorous assessment of the model's ability to generalize to new, unseen data.

3.2.Advanced Image Pre-processing Pipeline

To manage the variability inherent in dermatoscopic images and ready them for feature extraction, a structured, multi-step preprocessing pipeline was constructed. Initially, all images were resized to a consistent resolution of 224×224 pixels, ensuring uniform input dimensions for the models used in subsequent stages.

Hair removal was identified as a crucial step, given that hair can obscure important lesion details and negatively affect model performance. This was addressed through an inpainting approach. The image was first transformed to grayscale, after which the **Canny edge detector** was applied to capture the fine lines characteristic of hair.The Canny edge detector was chosen for its ability to detect thin, well-localized edges with minimal noise, making it especially effective for isolating hair strands without capturing unrelated skin textures or lesion boundaries as described in J. Canny[18]

These edges were then expanded using dilation to create a mask covering the hair strands. The regions defined by this mask were reconstructed using the **TELEA algorithm** via **cv2.inpaint**, allowing the removal of hair while preserving the skin's visual texture.The masked regions were reconstructed using the **TELEA inpainting algorithm** as implemented in OpenCV's cv2.inpaint, based on Telea's method [19] utilizing the fast marching technique. This algorithm was chosen for its **simplicity, speed, and its ability to preserve local texture** while seamlessly filling small occluded areas

In tandem, enhancement techniques were applied to emphasize key visual features. Two parallel methods were employed:

CLAHE (Contrast Limited Adaptive Histogram Equalization) was performed on the grayscale image, where contrast adjustments were made locally rather than globally. This allowed finer lesion textures and borders to be revealed without introducing excessive noise. CLAHE was selected based on Zuiderveld's formulation [20], as it effectively enhances local contrast while preventing over-amplification of noise—a limitation seen in traditional global histogram equalization. By operating on small contextual regions and applying contrast limiting, CLAHE preserves fine lesion textures and boundaries critical for accurate skin lesion analysis.

To improve model robustness and compensate for the limited dataset size, various data augmentation strategies were introduced. Cleaned images were subjected to random horizontal and vertical flips, 90-degree rotations, and subtle changes in brightness and contrast. These transformations preserved the clinical characteristics of the lesions while introducing variability, thereby enriching the training dataset and helping to mitigate overfitting.

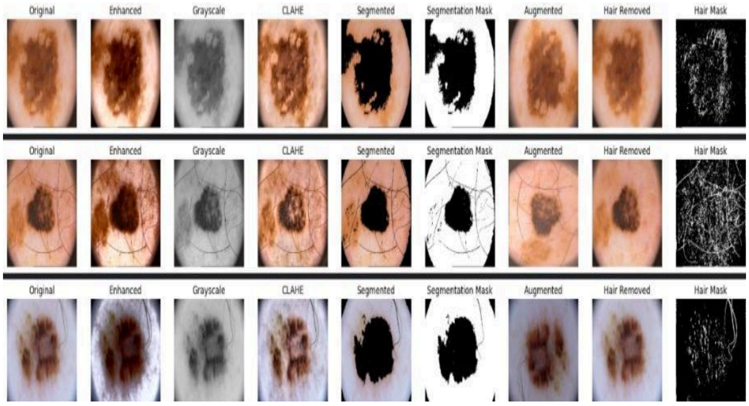


Fig 4. The image illustrates the key steps in our preprocessing pipeline for dermatoscopic skin lesion images. It includes the original image, sharpened and CLAHE-enhanced versions, grayscale conversion, segmentation mask, augmented image, and hair removal using inpainting.

3.3.Feature Extraction via Transfer Learning

The core of the feature extraction process relies on Transfer Learning, utilizing the MobileNetV2 architecture pre-trained on the ImageNet dataset. Instead of training a deep learning model from scratch, MobileNetV2 is employed as a fixed feature extractor. **MobileNetV2** [21] is a deep neural network architecture designed for efficiency in low-resource environments. Its improvements over previous versions are achieved through two main design principles:

Inverted Residuals:
Feature maps are first expanded using a 1×1 convolution, then processed with a depthwise separable 3×3 convolution, and finally reduced using a linear 1×1 convolution. This inverted structure allows important information to be retained while reducing model size.

Linear Bottlenecks:
Non-linear activations are applied only in the intermediate layers. The final projection uses a linear transformation without activation to prevent loss of essential features in low-dimensional space.

The model is instantiated without its final classification layer (include_top=False), and a Global Average Pooling layer is appended to its output. For each preprocessed input image, the model processes it through its convolutional layers and outputs a 1280-dimensional feature vector. This vector serves as a rich, high-level representation of the image's content, capturing complex patterns, textures, and shapes relevant to lesion classification.

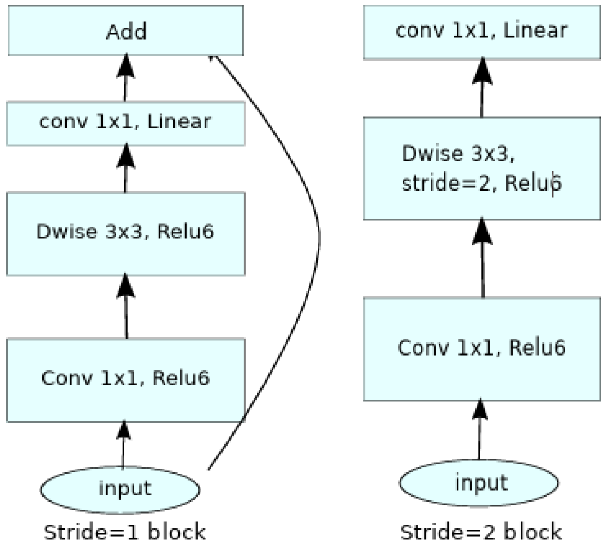


Fig 5. Architecture of MobileNetV2 building blocks as demonstrated in Sandle et al.[21]

3.4. Classifier Training and Comparative Analysis

The extracted feature vectors are used to train and evaluate two state-of-the-art gradient boosting models: XGBoost (Extreme Gradient Boosting) and LightGBM (Light Gradient Boosting Machine). XGBoost and LightGBM are two widely used gradient boosting frameworks known for their speed, accuracy, and scalability. XGBoost, introduced by Chen and Guestrin[22], is a tree-based ensemble learning method that uses second-order gradient information and includes optimizations like sparsity-aware split finding and parallel computation. It is especially effective on structured or tabular data and has become a go-to model for many machine learning competitions. LightGBM, developed by Ke et al.[23], builds upon similar principles but introduces techniques like Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to reduce memory usage and training time without compromising accuracy. LightGBM also grows trees leaf-wise rather than level-wise, which can lead to better accuracy but may risk overfitting on small datasets. In comparison, XGBoost tends to be more robust with small or noisy data, while LightGBM offers faster training and better scalability for large datasets. Both models can be effectively used in hybrid pipelines—such as combining CNN-based deep features with gradient boosting classifiers—to handle complex classification tasks like skin cancer detection.

Both models are trained on the features from the training set and evaluated against the validation set. An early stopping mechanism is employed, monitoring the multi-class log loss on the validation data to prevent overfitting and determine the optimal number of boosting rounds. After training, the final models are evaluated on the unseen test set to provide an unbiased assessment of their generalization performance.

3.5.Results

Evaluation of classifier performance was conducted using several established metrics. **Accuracy**, defined as the proportion of correctly classified instances, offers a fundamental measure of overall performance. However, as highlighted in performance evaluation studies, accuracy may be misleading in the presence of class imbalance. To provide a more nuanced analysis, **precision**, **recall**, and **F$_1$-score** were calculated per class and aggregated using **macro**, **micro**, and **weighted averages**. The **macro-average** treats each class equally by computing the metric independently for each class and then averaging, while the **micro-average** pools contributions across classes, thus reflecting performance weighted by class prevalence.

To account for chance agreement in classification tasks, **Cohen's Kappa score** was employed as a robust alternative to raw accuracy. Kappa measures the degree of agreement between predicted and actual labels beyond expected random agreement, making it particularly valuable in imbalanced or multi-class settings. In addition, confusion matrices were generated to visualize the distribution of true positives, false positives, true negatives, and false negatives across classes, facilitating detailed error analysis.

Table 2. ***Performance of XGBoost and LightGBM classifiers on the test dataset using standard evaluation metrics.***

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1 Score (Macro)	Kappa Score	Mean AUC
XGBoost (Test)	0.854846	0.835417	0.805782	0.817977	0.828480	0.981891
LightGBM (Test)	0.850826	0.830225	0.798755	0.810408	0.823732	0.980359

4. Conclusion

Based on the evaluation results and comparative analysis, our proposed hybrid framework has demonstrated reliable and efficient performance in multi-class skin cancer classification. By combining deep learning-based feature extraction through MobileNetV2 with robust machine learning classifiers like XGBoost and LightGBM, the system effectively capitalizes on the strengths of both paradigms. The convolutional backbone captures rich spatial features from dermatoscopic images, while gradient boosting classifiers provide accurate, interpretable decision boundaries. This modularity also offers flexibility for deployment in diverse clinical or low-resource settings.

From the performance metrics summarized in the table, XGBoost slightly outperforms LightGBM across all key evaluation parameters. XGBoost achieves an accuracy of 85.48%, a macro F1-score of 0.8179, and a Kappa Score of 0.8284—suggesting strong generalization across all lesion classes, including the underrepresented ones. LightGBM closely follows with a respectable accuracy of 85.08% and macro F1-score of 0.8104, making it a competitive alternative with lower training time and computational cost. These metrics, coupled with the confusion matrices, show that both models perform well at classifying common lesions like pigmented benign keratosis and basal cell carcinoma, while classes like seborrheic keratosis and actinic keratosis remain more challenging, partly due to visual similarities and class imbalance.

Overall, the results validate the efficacy of our hybrid learning pipeline for automated skin cancer diagnosis. The integration of preprocessing techniques like hair removal, contrast enhancement, and augmentation ensured high-quality inputs, while transfer learning via MobileNetV2 enabled efficient representation learning. By using ensemble classifiers instead of conventional fully connected layers, the model achieved superior classification performance without incurring significant computational overhead. This approach not only offers diagnostic support to dermatologists but also paves the way for scalable, real-world deployment of skin lesion analysis tools in primary care or remote settings.

5. References

1. Ioannou, Y., Robertson, D., Zikic, D., Kontschieder, P., Shotton, J., Brown, M., Criminisi, A.: Decision Forests, Convolutional Networks and the Models in-Between. MSR Technical Report MSR-TR-2015-58, Microsoft Research (2016).

2. Afza, F., Sharif, M., Khan, M.A., Tariq, U., Yong, H.-S., Cha, J.: Multiclass Skin Lesion Classification Using Hybrid Deep Features Selection and Extreme Learning Machine. Sensors 22(3), 799 (2022).

3. Hasan, M.Z., Rifat, F.Y.: Hybrid Ensemble of Segmentation-Assisted Classification and GBDT for Skin Cancer Detection with Engineered Metadata and Synthetic Lesions. arXiv preprint arXiv:2506.03420 (2024).

4. Ali, R., Ragb, H.K.: Skin Lesion Segmentation and Classification Using Deep Learning and Handcrafted Features. arXiv preprint arXiv:2112.10307 (2021).

5. Kaur, R., GholamHosseini, H., Lindén, M.: Advanced Deep Learning Models for Melanoma Diagnosis in Computer-Aided Skin Cancer Detection. Sensors 2025, 25, 594.

6. Houssein, E.H., Abdelkareem, D.A., Hu, G., Abdel Hameed, M., Ibrahim, I.A., Younan, M.: An effective multiclass skin cancer classification approach based on deep convolutional neural network. Cluster Computing (2024) 27:12799–12819.

7. Milton, M.A.A.: Automated Skin Lesion Classification Using Ensemble of Deep Neural Networks in ISIC 2018 Challenge. arXiv preprint arXiv:1901.10802 (2019).

8. Berseth, M.: ISIC 2017 – Skin Lesion Analysis Towards Melanoma Detection. arXiv preprint arXiv:1703.00523 (2017).

9. Joseph, S., Olugbara, O.O.: Preprocessing Effects on Performance of Skin Lesion Saliency Segmentation. *Diagnostics* **12**(2), 344 (2022).

10. Naqvi, M., Gilani, S.Q., Syed, T., Marques, O., Kim, H.-C.: Skin Cancer Detection Using Deep Learning—A Review. *Diagnostics* **13**(11), 1911 (2023).

11. Lin, B.S., Michael, K., Kalra, S., Tizhoosh, H.R.: Skin Lesion Segmentation: U-Nets versus Clustering. *To appear in IEEE SSCI 2017*, Honolulu, Hawaii, USA (2017).

12. Perez, F., Vasconcelos, C., Avila, S., Valle, E.: Data Augmentation for Skin Lesion Analysis. *arXiv preprint arXiv:1809.01442* (2018).

13. Mirikharaji, Z., Abhishek, K., Bissoto, A., Barata, C., Avila, S., Valle, E., Celebi, M.E., Hamarneh, G.: A Survey on Deep Learning for Skin Lesion Segmentation. *arXiv preprint arXiv:2206.00356* (2023).

14. nodoubttome: Skin Cancer: 9 Classes-ISIC. Kaggle (2023). https://www.kaggle.com/datasets/nodoubttome/skin-cancer9-classesisic, last accessed 2025/07/27.

15. Codella, N., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., Halpern, A.: Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC). arXiv:1710.05006 [cs.CV] (2017).

16. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., Halpern, A.: Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). arXiv:1902.03368 [cs.CV] (2019).

17. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5, 180161 (2018).

18. J. Canny, "A Computational Approach to Edge Detection," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679-698, Nov. 1986, doi: 10.1109/TPAMI.1986.4767851

19. A. Telea, “An Image Inpainting Technique Based on the Fast Marching Method,” *Journal of Graphics Tools*, vol. 9, no. 1, pp. 23–34, 2004.

20. Zuiderveld, K.: Contrast Limited Adaptive Histogram Equalization. In: Heckbert, P.S. (ed.) *Graphics Gems IV*, pp. 474–485. Academic Press, San Diego (1994).

21. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: Inverted Residuals and Linear Bottlenecks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520. IEEE, Salt Lake City (2018).

22. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM, New York (2016). https://doi.org/10.1145/2939672.2939785

23. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 3146–3154 (2017).