Phase-I Evaluation

Advanced Ensemble Deep Learning Architecture for Automated Skin Cancer Classification: A Comprehensive EfficientNet-B3 and InceptionV3 Hybrid Framework for Multi-Class Dermoscopic Image Analysis

Presented by

Amit Kumar Ghosh
ID :221-15-4650

Department of CSE
Daffodil International University

Supervised by

Mr. Md Assaduzzaman
Senior Lecturer
Department Of **CSE**
Daffodil International University

Outline

- Introduction
- Problem Statement
- Motivation
- Objective
- Related Works
- Comparison Between Existing Works
- Gap Analysis
- Proposed Methodology
- Q & A

Introduction

Skin cancer ranks among the most prevalent and life-threatening malignancies worldwide, with incidence continuing to rise. Early detection through dermoscopic imaging can dramatically improve patient outcomes, yet manual interpretation remains time-consuming and variable across clinicians. Integrating advanced AI models into the diagnostic workflow holds the promise of consistent, high-accuracy screening. In this work, we present an efficient ensemble of EfficientNet-B3 and InceptionV3, tailored to the HAM10000 dataset, combining robust feature extraction and optimized training strategies to enhance multiclass lesion classification while maintaining inference speed and resource efficiency.

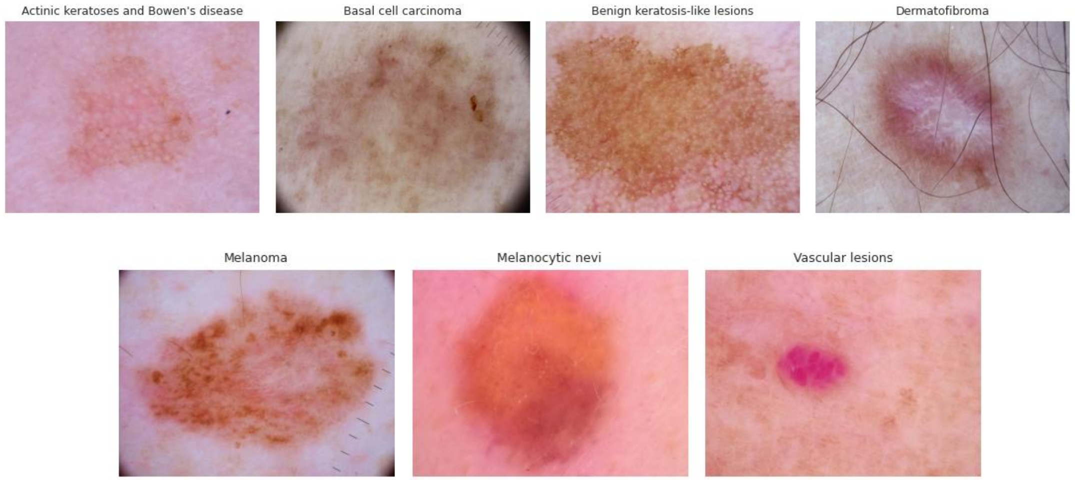


Fig: Sample Images of 7 Classes

Problem Statement

- Manual dermoscopic diagnosis is **time-consuming**, slowing clinical workflows.
- Interpretation varies significantly between clinicians, leading to inconsistent results.
- Extreme **class imbalance** in HAM10000 (e.g., rare lesion types) reduces model recall on minority classes.
- Single CNN models often overfit abundant classes while under-detecting critical but scarce lesions.
- Many existing ensembles improve accuracy but incur high inference latency and resource use.

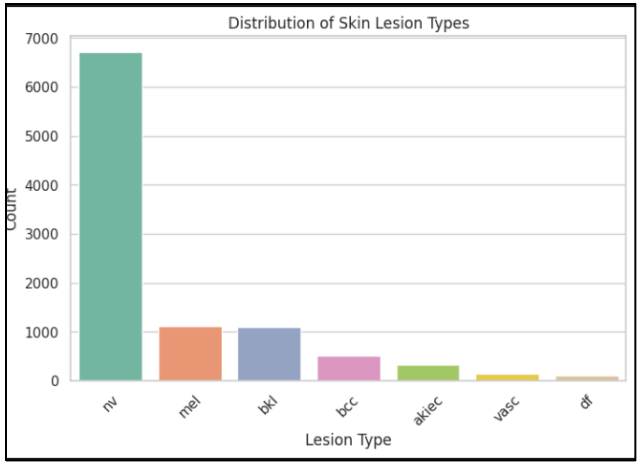


Fig: Severe Class Imbalance

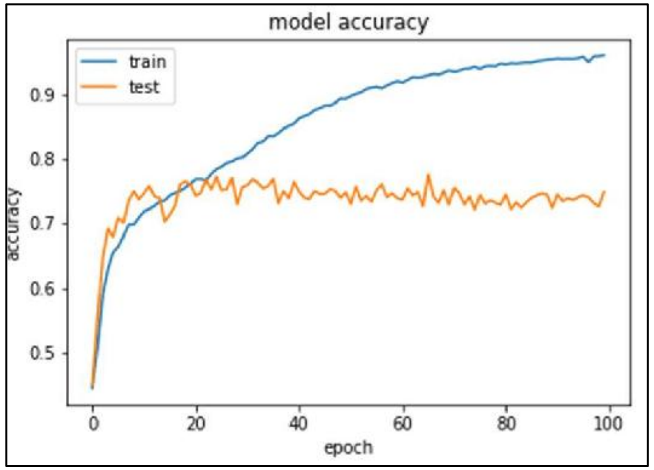


Fig: Overfitting in Single CNN Models

Motivation

- **Address High Mortality:** Skin cancer remains one of the deadliest malignancies when diagnosed late. Leveraging deep learning can improve early detection, directly reducing mortality rates.
- **Reduce Diagnostic Variability:** Manual dermoscopic interpretation suffers from inter-observer differences. An AI-driven ensemble ensures consistent lesion analysis across diverse clinical settings.
- **Mitigate Class Imbalance:** Rare lesion types in HAM10000 hinder traditional CNN performance. Oversampling and mixed-precision training help the ensemble learn underrepresented classes more effectively.
- **Enhance Clinical Efficiency:** Real-time inference with EfficientNet-B3 and InceptionV3 ensemble minimizes latency, allowing seamless integration into busy workflows without sacrificing accuracy.
- **Bridge Research Gaps:** Few studies combine lesion-level splitting, advanced augmentation, and hybrid architectures for skin-lesion classification. This work fills that gap by unifying best practices in a single, optimized pipeline.

Objectives

- Develop a **hybrid ensemble model** combining EfficientNet-B3 and InceptionV3 for robust multiclass classification on the HAM10000 dataset.
- Implement lesion-level train/validation/test splitting to prevent data leakage and ensure genuine performance estimates.
- Apply real-time data augmentation (flips, rotations, color jitter) and targeted oversampling to mitigate extreme class imbalance.
- Leverage mixed-precision (FP16) training and gradient accumulation to enable larger effective batch sizes while optimizing memory and speed.
- Design an ensemble strategy that balances accuracy and inference efficiency through weighted averaging of softmax outputs.
- Integrate checkpointing based on validation metrics to retain and deploy the optimal model configuration.

Related Works

- M. Shakya et al. Developed three deep learning–machine learning hybrids for melanoma on ISIC 2018, using active-contour segmentation plus scaling, denoising, and enhancement. Their top model fused **ResNet-18 and MobileNet_V2 features with an SVM classifier**, reaching 92.87% accuracy. Generalization may be limited by reliance on a single dataset.
- P. Georgiadis et al. Created the Data Merger App to unify multiple skin-lesion datasets into large "hyperdatasets." Evaluated VGG16, ResNet50, MobileNetV3-small, DenseNet-161, and ViT: achieved up to 91.87% accuracy on 9-way classification with ViT, but accuracy fell to 58% for 32 classes, underscoring scaling challenges.
- M. Abdel et al. Introduced AEDHOA, a metaheuristic combining SRIS, ELCS, APS, and DES for robust feature selection. Tested on CEC benchmarks, UCI sets, and a skin-cancer dataset; achieved accuracies from 76% to 100%. Effectiveness on high-dimensional data is strong, though computational cost grows with dataset size.
- Rodrigue et al. Compared YOLO v7 (transfer-learned) and a custom CNN on 2,792 augmented images for Basal Cell Carcinoma, Squamous Cell Carcinoma, and Melanoma. Their CNN attained 90.12% accuracy, 85.55% sensitivity, and 92.57% specificity, though limited sample size may constrain wider applicability.
- A. A. Hussein et al. Proposed a hybrid quantum CNN + BiLSTM + MobileNetV2 model for skin cancer classification at 32×32 and 128×128 resolutions. Reported 89.3% accuracy, 89.81% F1, and 94.33% recall on malignant lesions. High performance is tempered by system complexity and quantum-component requirements.

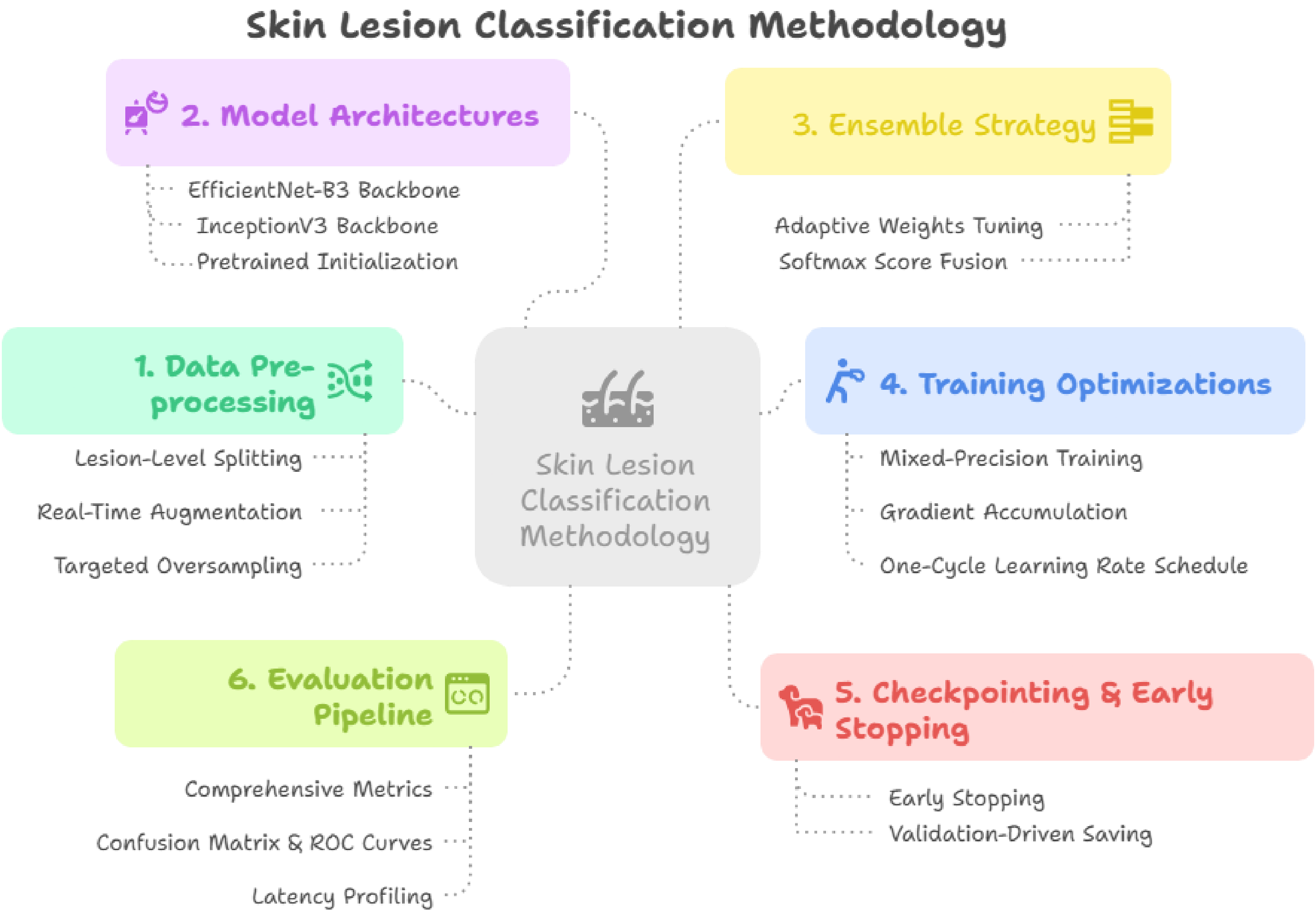
Comparison Between Existing Works

Author and Year	Used Model	Achieved Accuracy	Key Contribution
M. Shakya et al.(2025)	ResNet-18 + MobileNet V2 features fed to SVM	92.87%	Demonstrated that a hybrid deep-learning/ML pipeline outperforms single CNNs on ISIC-2018 while remaining lightweight.
P. Georgiadis et al.(2025)	Visual Transformer (ViT) on merged “hyper-datasets”	91.87% (9 classes) / 58% (32 classes)	Introduced Data Merger App to automatically combine disparate skin-lesion datasets and showed benefits—and limits—of large, diverse training sets.
M. Abdel et al.(2025)	AEDHOA-selected feature subset with classical classifier	0.76–1.00 (dataset-dependent)	Proposed Adaptive Enhanced Diversified Hiking Optimization Algorithm for robust feature selection on high-dimensional skin-cancer data.
Rodrigue et al.(2025)	Custom CNN (compared with YOLO v7 TL)	90.12%	Built a compact CNN that slightly outperformed YOLO v7 for three common cancers using only 2,792 augmented images.
A. A. Hussein .(2025)	Hybrid Quantum CNN + BiLSTM + MobileNet V2	89.3%	Showed quantum-inspired feature extraction plus temporal context can boost lesion recognition, though with higher system complexity.
Our Proposed Model(2025)	**EfficientNet-B3 + InceptionV3** weighted-softmax ensemble	**93% (HAM10000 validation)**	Combines **lesion-level split**, targeted oversampling, mixed-precision training, and **dual-backbone ensemble** to improve minority-class recall without adding inference lag.

Gap Analysis

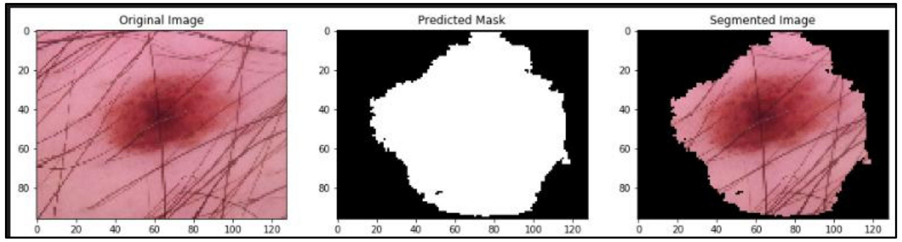
- **Lack of Multi-Dataset Generalization:** Training is limited to HAM10000, missing variations in imaging, demographics, or lesions from real-world sources, reducing transferability to diverse environments like varying skin tones in Bangladesh.
  Future Solution: Integrate ISIC 2018/2019 or PAD-UFES-20 for cross-dataset training and fine-tuning, targeting 85%+ accuracy on unseen data via transfer learning.
- **Absence of Real-World Deployment Testing:** Inference efficiency is optimized but unprofiled on clinic hardware, leaving latency claims theoretical and overlooking low-resource bottlenecks in rural settings.
  Future Solution: Benchmark on NVIDIA Jetson or mobile CPUs, apply quantization for sub-100ms times, and pilot a prototype app for clinician input.
- **Limited Explainability and Interpretability:** The model lacks tools to explain predictions, essential for clinician trust in medical black-box systems.
  Future Solution: Add Grad-CAM or SHAP for heatmaps, integrate into evaluation, and validate with dermatologist reviews for clinical alignment.
- **Incomplete Handling of Edge Cases and Robustness:** Augmentation addresses imbalance, but no stress-testing against noise, artifacts, or adversarial examples; per-class metrics are unreported from runs.
  Future Solution: Inject synthetic noise and adversarial training; conduct ablation studies to boost minority F1 above 90% via hyperparameter tuning.
- **Scalability to More Classes or Modalities:** Tailored to 7 classes, it doesn't scale to finer sub-types or multi-modal data like patient metadata (age, UV exposure).

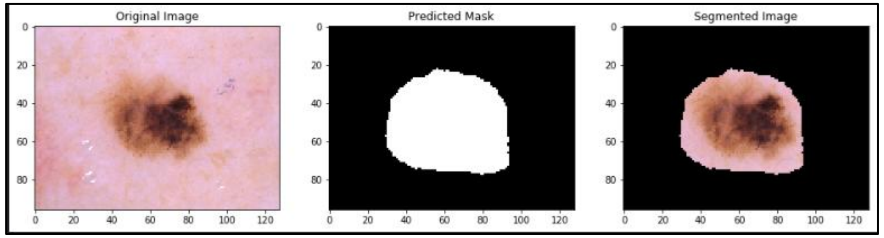
Proposed Methodology



Proposed Methodology

- Data Preprocessing
  - **Lesion-Level Splitting:** Partition the HAM10000 dataset by unique lesion identifiers to ensure no images of the same lesion appear across train, validation, and test sets, eliminating data leakage.
  - **Real-Time Augmentation:** Apply on-the-fly transformations (random flips, rotations up to ±30°, color jitter, and random crops) during training to increase data diversity without inflating storage.





  - **Targeted Oversampling:** Dynamically replicate underrepresented classes in each epoch to balance batch composition and improve minority-class learning.
- Model Architectures
  - **EfficientNet-B3 Backbone:** Leverage EfficientNet's compound scaling to extract robust features with relatively low parameter count and FLOPs.
  - **InceptionV3 Backbone:** Capture multi-scale spatial patterns via its parallel convolutional branches, complementing EfficientNet features.
  - **Pretrained Initialization:** Initialize both networks with ImageNet weights to accelerate convergence and leverage transferable representations.
- Ensemble Strategy
  - **Softmax Score Fusion:** For each input, compute class probabilities from both backbones, then combine via weighted averaging (e.g., 0.6·EfficientNet + 0.4·Inception) to balance their strengths.
  - **Adaptive Weights Tuning:** Optimize fusion weights on the validation set using grid search to maximize macro-F1 score, emphasizing minority classes.

Proposed Methodology

- Training Optimizations
  - **Mixed-Precision Training:** Use FP16 arithmetic with dynamic loss scaling to accelerate training and reduce GPU memory usage without sacrificing numerical stability.
  - **Gradient Accumulation:** Accumulate gradients over multiple mini-batches to simulate larger batch sizes, stabilizing updates when hardware memory is limited.
  - **One-Cycle Learning Rate Schedule:** Employ a triangular learning-rate policy that ramps up then down over each epoch to foster faster convergence and avoid local minima.
- Checkpointing & Early Stopping
  - **Validation-Driven Saving:** After each epoch, save model weights when macro-F1 on the validation set improves, ensuring the best ensemble components are retained.
  - **Early Stopping:** Halt training if no validation F1 improvement occurs for 10 consecutive epochs to prevent overfitting and conserve resources.
- Evaluation Pipeline
  - **Comprehensive Metrics:** Report overall accuracy, per-class precision, recall, and F1—highlighting performance on rare lesion categories.
  - **Confusion Matrix & ROC Curves:** Generate normalized confusion matrices and per-class ROC curves to visualize error patterns and diagnostic thresholds.
  - **Latency Profiling:** Measure end-to-end inference time on target hardware to confirm that the ensemble meets clinical throughput requirements.

Any Question

THANK YOU