SKIN CANCER DETECTION USING HYBRID CNN–VISION TRANSFORMER ARCHITECTURE

Advanced Artificial Intelligence
Assignment 3 (Project Report)

Submitted By: 25k-7627 (Sanjinee)

1. Problem Statement

Automated skin cancer detection faces three critical challenges that hinder clinical deployment:

Challenge 1: Severe Class Imbalance
Medical datasets like ISIC 2018 exhibit pronounced imbalance where benign lesions (particularly melanocytic nevus) vastly outnumber malignant cases. This creates a risk where models achieve high overall accuracy by correctly classifying abundant benign cases while failing to detect rare but clinically critical malignancies like melanoma.

Challenge 2: Limited Training Data
Unlike natural image datasets with millions of samples, medical imaging datasets are constrained by the cost and expertise required for professional annotation. The ISIC 2018 dataset contains only 10,015 images across seven diagnostic classes, making it challenging to train deep learning models that typically require large-scale data.

Challenge 3: Lack of Clinical Interpretability
Pure deep learning models operate as "black boxes," providing predictions without transparent reasoning. For clinical deployment, dermatologists require visual evidence showing which image regions influenced the diagnosis, ensuring the model focuses on clinically relevant features rather than spurious correlations.

Research Objective:
Develop a hybrid CNN-Vision Transformer architecture that achieves balanced performance across all diagnostic classes (prioritizing Macro F1 Score over raw accuracy) while providing explainable predictions suitable for clinical decision support.

---

2. Base Paper Results vs. Documented Results

Base Paper: AViT (Zhang et al., 2025)

Focus: Lesion segmentation (pixel-level boundary delineation)
Dataset: HAM10000
Key Innovation: Domain-adapted Vision Transformers for small medical datasets
Performance: 96.3% accuracy on classification task
Methodology: Diagnosis-guided integration of CNN features with domain-adapted ViTs

Our Implementation: Hybrid CNN-ViT for Classification

Focus: Multi-class lesion classification (diagnostic categorization)
Dataset: ISIC 2018 (7 classes)
Key Innovation: Three-fold contribution:

1. Dual-strategy class imbalance mitigation (Weighted Random Sampling + Weighted Cross-Entropy Loss)
2. Domain-aware feature integration using acquisition metadata
3. Comprehensive explainability framework (Grad-CAM + SHAP-based patch attribution)

Performance Comparison:

Metric	Base Paper (AViT)	Our Implementation
Primary Task	**Segmentation**	**Classification**
Dataset	**HAM10000**	**ISIC 2018**
Accuracy	**96.3%**	**80.98%**
Macro F1 Score	**Not reported**	**0.7751**
Optimization Target	**Overall accuracy**	**Macro F1 (balanced performance)**
Explainability	**Not reported**	**Grad-CAM + SHAP analysis**

Critical Distinction:
While the base paper achieved higher raw accuracy, it did not report Macro F1 Score or address class imbalance explicitly. Our work prioritizes balanced performance across all classes, ensuring reliable detection of rare malignancies rather than maximizing overall accuracy through majority class predictions.

3. Methodology and Achievements

3.1 Dataset and Preprocessing

Dataset: ISIC 2018 Skin Lesion Analysis Challenge

- **Size:** 10,015 dermatoscopic images
- **Classes:** 7 diagnostic categories (MEL, NV, BCC, AKIEC, BKL, DF, VASC)
- **Challenge:** Severe imbalance (NV dominates; MEL, DF, VASC underrepresented)

Preprocessing Pipeline:

1. Image standardization to 224×224 pixels
2. Data augmentation: random rotations, horizontal/vertical flips, color jittering
3. Normalization using ImageNet statistics
4. Domain label encoding from metadata (acquisition source, clinical center)

3.2 Dual-Strategy Class Imbalance Mitigation

Strategy 1: Weighted Random Sampling
Computed inverse frequency weights for each class based on sample distribution. Applied weighted random sampler during training to ensure minority classes appear proportionally more often in training batches.

Strategy 2: Weighted Cross-Entropy Loss
Applied class-specific loss weights to impose heavier penalties on misclassifications of underrepresented classes during backpropagation.

Optimization Metric: Macro F1 Score (unweighted average of per-class F1 scores) to ensure no single class dominates performance evaluation.

3.3 Hybrid CNN-ViT Architecture

Component 1: CNN Feature Extractor (ResNet18)

- Pretrained on ImageNet for transfer learning
- Extracts fine-grained textural and structural information
- Captures local patterns: lesion borders, pigmentation networks, surface texture

Component 2: Domain-Adapted Vision Transformer

- 8 transformer layers (balanced for small dataset)
- Self-attention mechanism for global context modeling
- Domain adaptation module integrating acquisition source metadata
- Enables learning of source-specific biases while maintaining generalization

Architecture Flow:

Input Image (224×224) → ResNet18 Backbone (local features) → Feature Maps → Flattened & Projected → Token Sequences → Domain-Adapted ViT (global attention + domain labels) → Classification Head → 7-class prediction

3.4 Training Protocol

Framework: Skorch (scikit-learn compatible PyTorch wrapper)
Epochs: 40 (optimal checkpoint at Epoch 32)
Batch Processing: Weighted random sampling ensures balanced class representation
Loss Function: Weighted cross-entropy with inverse frequency weights
Validation Strategy: Held-out validation set with Macro F1 Score monitoring

3.5 Explainability Framework

Technique 1: Gradient-weighted Class Activation Mapping (Grad-CAM)

- Heatmaps consistently localized attention on clinically relevant structures
- Model successfully identified lesion borders, asymmetry patterns, and pigmentation networks
- Visual evidence confirms integration of local CNN features with global ViT context

Technique 2: SHAP-based Patch Attribution

- Positive contributions (blue/green patches): Features confirming predicted diagnosis
- Negative contributions (red patches): Features suggesting alternative diagnoses
- Internal conflict resolution: Model weighs competing evidence before final prediction
- Complex cases show high negative contributions successfully overridden by stronger confirming evidence

Key Finding: Model demonstrates internal conflict resolution in ambiguous cases, mirroring expert dermatologist diagnostic reasoning by weighing conflicting visual evidence before final classification.

3.6 Key Achievements

1. **Balanced Performance:** Achieved 0.7751 Macro F1 Score, ensuring reliable detection across all seven classes including rare malignancies
2. **Domain Generalization:** Successfully integrated acquisition metadata to adapt to diverse clinical settings and imaging equipment
3. **Clinical Interpretability:** Provided visual and quantitative evidence of diagnostic reasoning through dual explainability techniques
4. **Architectural Efficiency:** Designed for small medical datasets (10K images) while avoiding overfitting through domain adaptation and careful layer sizing
5. **Transparent Conflict Resolution:** Demonstrated model's ability to handle visually ambiguous cases by quantifying internal feature conflicts
6. **Primary Metric: Validation Macro F1 Score = 0.7751**
   This score represents unweighted average performance across all seven diagnostic classes, ensuring minority classes receive equal consideration in evaluation.

7. **Secondary Metric: Validation Accuracy = 80.98%**
   While lower than base paper's 96.3%, this reflects our deliberate optimization for balanced performance rather than maximizing overall accuracy through majority class predictions.

Future Work

- **Model Enhancements** - Technical improvements to the architecture
- **Clinical Deployment** - Practical implementation considerations
- **Interpretability and Collaboration** - Explainability and broader applications
- **Multi-Task Learning:** Combine segmentation and classification tasks to leverage synergies
- **Advanced Augmentation:** Implement CutMix and MixUp for improved minority class representation
- **Ensemble Methods:** Integrate multiple hybrid architectures (ViT-DenseNet, ViT-EfficientNet) for robustness
- **Cross-Attention:** Refine feature integration between CNN maps and ViT tokens
- **Expanded Metadata:** Include patient demographics (age, skin type, lesion location) for personalized assessment