



**Project Report**

**Summer Internship (UNT)**

**Health Informatics**

**Agent-Based AI System for Melanoma Classification**

**BY**

**SURAM JAHNAVI REDDY**

**(229X1A3392)**

**G. PULLA REDDY ENGINEERING COLLEGE (AUTONOMOUS) KURNOOL, ANDHRA PRADESH**

**CONTENTS**

|  |
| --- |
| **Chapter 1:** Abstract  **Chapter 2:** Introduction  **Chapter 3:** Materials and Methods  **Chapter 4:** Experimental Design and Workflow  **Chapter 5:** Results  **Chapter 6:** Discussion  **Chapter 7:** Conclusion  **Chapter 8:** Acknowledgement |

**Agent-Based AI System for Melanoma Classification**

# **Abstract**

**Background:**Melanoma is a deadly form of skin cancer requiring early detection. While dermoscopic imaging aids diagnosis, manual interpretation is time-consuming and subjective. Deep learning offers a promising solution for automated lesion classification.

**Objective:**This project aims to build a modular, agent-based AI system to classify dermoscopic images as melanoma or benign, combining visual and language models for accuracy and interpretability.

**Methodology:**Inspired by the PathFinder framework, the system uses:

* A Triage Agent (ViT model) to assign malignancy scores and filter high-risk lesions.
* Optional Navigation Agent to direct decision flow.
* Decision Agents (CLIP, GIT) to generate explanations and malignancy estimates.

A total of 150 images from the HAM10000 dataset were processed through the pipeline. Evaluation metrics included True/False Positives/Negatives, Sensitivity, and Specificity across triage thresholds (0.4, 0.5, 0.6). Two systems were compared:

* S1: ViT-only
* S2: ViT triage + CLIP/GIT

**Results:**System S2 provided interpretability via natural language outputs but underperformed in sensitivity and accuracy compared to S1. CLIP and GIT often gave incorrect outputs due to lack of domain-specific training.

**Conclusion:**While the agent-based approach offers a novel framework, results emphasize the need for fine-tuning general-purpose models on medical data. The study underscores the importance of domain adaptation for effective and trustworthy AI diagnostics.

**Introduction**Skin cancer is one of the most common forms of cancer globally, with melanoma being its most aggressive and life-threatening variant. Early detection and accurate classification of melanoma are critical for improving patient outcomes. Dermoscopic imaging has significantly enhanced clinicians’ ability to identify malignant lesions, but manual examination is time-consuming, prone to subjectivity, and often limited by the expertise of the observer. To address these challenges, artificial intelligence (AI), particularly deep learning, has emerged as a powerful tool in the field of medical image analysis.

Transformer-based models such as Vision Transformers (ViT), CLIP, and GIT have demonstrated strong performance in computer vision and vision-language tasks. These models can automatically learn rich features from large-scale image datasets and provide predictions that often rival human-level accuracy. However, many existing AI systems act as black boxes, offering little insight into the reasoning behind their predictions—limiting their practical applicability in high-stakes domains like healthcare.

To overcome this limitation, this project proposes a **multi-agent AI system** that combines classification and interpretability in a unified pipeline. Inspired by the **PathFinder framework**, the system divides the diagnostic process into modular agents—each responsible for a specific task such as lesion screening, visual-textual analysis, and decision justification. This agent-based design enhances transparency, allows step-wise reasoning, and facilitates collaboration between AI components, mirroring the workflow of human clinical decision-making.

**Materials and Methods**

**Dataset:**  
Images from ISIC and HAM10000 datasets (~5,000+ total, 150 used in initial experiments). Lesions labeled melanoma vs benign.Ground truth derived from HAM10000\_metadata.csv (dx column).

**System Design:**  
Triage Agent: ViT model (actavkid/vit-large-patch32-384-finetuned-skin-lesion-classification) computes malignancy scores.

**Decision Agents:**  
CLIP: Performs image-text classification using textual prompts (e.g., "malignant", "benign").  
GIT: Generates captions; heuristic rules extract malignancy scores from generated descriptions.

**Malignancy Score Computation:**  
ViT: Combined softmax probabilities of melanoma, BCC, and AKIEC.  
CLIP/GIT: Numeric probability and explanation extracted from text output.

**Evaluation Metrics:**  
Classification accuracy, true/false positives and negatives.  
Sensitivity, specificity at triage thresholds (0.4, 0.5, 0.6).

**Experimental Design and Workflow**

The project followed a phased experimental approach over a six-week internship period:

* **Week 1: Initial Model Exploration:**
  + **Objective:** Identify and evaluate a suitable pre-trained deep learning model for dermoscopic image classification using transformer-based architectures.
  + **Activities:** Explored Hugging Face models, selected actavkid/vit-large-patch32-384-finetuned-skin-lesion-classification. Downloaded 150 HAM10000 images and developed a basic classification pipeline in Google Colab for loading, preprocessing, inference, and display of predictions.
* **Week 2: Model Validation and Performance Metrics:**
  + **Objective:** Validate the classification model using 10-fold cross-validation and evaluate its performance using Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) metrics.
  + **Activities:** Mapped image filenames with metadata, encoded labels to binary (melanoma vs. non-melanoma). Implemented stratified 10-fold cross-validation and computed AUC for each fold using sklearn.metrics.roc\_auc\_score. Plotted an overall ROC curve.
* **Week 3: Multi-Agent Communication Simulation:**
  + **Objective:** Implement and simulate multi-agent communication between AI models for lesion analysis, drawing inspiration from the PathFinder paper.
  + **Activities:** Developed and integrated preprocessing\_agent, border\_detection\_agent, triage\_agent, and navigation\_agent. Evaluated the integrated full\_agent\_pipeline across 150 images. Explored candidate AI models (BLIP-2, GIT) for generating verbal descriptions of lesions.
* **Week 4: Vision-to-Text (V2T) Model Integration:**
  + **Objective:** Integrate V2T models (BLIP-2, GIT) with the ViT classifier and analyze their outputs on dermoscopic images to explore multi-model fusion.
  + **Activities:** Developed a standardized V2T prompt ("What is the probability of the given image being malignant melanoma?"). Integrated BLIP-2 and GIT into the classification pipeline. Processed 10 sample images, compiled outputs (ViT, BLIP2, GIT malignancy scores) into a unified CSV, and measured processing time and system utilization on Google Colab Pro.
* **Week 5: Agent Communication and V2T Evaluation:**
  + **Objective:** Simulate agent communication between the Triage agent (ViT) and verbal description agents (CLIP and GIT), comparing their outputs against ground truth lesion labels.
  + **Activities:** Processed 150 HAM10000 images. Ran ViT for malignancy scores and triage. Integrated CLIP (prompt-based classification) and GIT (descriptive captions). Mapped ground truth labels and logged comprehensive results (ImageID, GroundTruthLabel, VIT\_Malignancy\_Score, VIT\_Malignant, Triage, CLIP\_Answer, CLIP\_Malignant, GIT\_Answer, GIT\_Malignant) in a structured CSV.
* **Week 6: Performance Evaluation and Disagreement Resolution:**
  + **Objective:** Evaluate the performance of two decision systems for melanoma classification, perform comparative analysis using key metrics, visualize performance across varying triage thresholds, and propose a resolution strategy for model disagreement.
  + **Activities:** Implemented and compared two classification systems: S1 (ViT-only) and S2 (multi-agent pipeline with ViT for triage and CLIP/GIT for final decision). Calculated TP, TN, FP, FN, Sensitivity, and Specificity for both systems at thresholds 0.4, 0.5, and 0.6.

**Results**

The project successfully implemented a modular, agent-based AI system for classifying dermoscopic images. However, results showed significant performance gaps, especially with the multi-agent system (System S2).

* **ViT Model (System S1)**:  
  Achieved consistent performance across all thresholds, with high specificity (≥0.92) and moderate sensitivity (~0.59). The ROC and AUC from 10-fold cross-validation supported ViT’s reliability for melanoma detection.
* **CLIP and GIT Integration (System S2)**:  
  Despite the design goal of interpretability, the decision agents frequently failed to detect malignant lesions. Sensitivity dropped drastically (as low as 0.037), indicating poor recall for melanoma. Specificity was slightly improved but at the cost of missing critical cases.
* **Threshold Analysis**:  
  Threshold tuning did not significantly improve the multi-agent system's ability to balance sensitivity and specificity. The logical OR resolution strategy for CLIP and GIT decisions was conservative but failed to recover true positives.

**System S1 (ViT-Only)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Triage Threshold** | **TP** | **TN** | **FP** | **FN** | **Sensitivity** | **Specificity** | | --- | --- | --- | --- | --- | --- | --- | | **0.4** | 16 | 114 | 9 | 11 | 0.593 | 0.927 | | **0.5** | 16 | 114 | 9 | 11 | 0.593 | 0.927 | | **0.6** | 16 | 114 | 9 | 11 | 0.593 | 0.927 |   **System S2 (ViT + CLIP/GIT)**   | **Triage Threshold** | **TP** |  | **TN** | **FP** | **FN** | **Sensitivity** | **Specificity** | | --- | --- | --- | --- | --- | --- | --- | --- | | **0.4** | 2 |  | 117 | 6 | 25 | 0.074 | 0.951 | | **0.5** | 1 |  | 117 | 6 | 26 | 0.037 | 0.951 | | **0.6** | 2 |  | 118 | 5 | 26 | 0.037 | 0.959 | |

# **Discussion**

The experimental results of this project indicate that the **agent-based system (System S2)** did not perform as well as initially anticipated. While the motivation behind introducing multiple agents—especially CLIP and GIT for reasoning and interpretability—was to improve diagnostic support, the actual performance metrics (sensitivity, specificity, and overall classification accuracy) were **inferior to the baseline ViT-only system (System S1)**. In particular, **System S2 misclassified more lesions**, leading to increased false positives and false negatives across various triage thresholds.

One of the key limitations was the use of **general-purpose vision-language models (CLIP and GIT)** that were not trained or fine-tuned on medical data. As a result, their interpretations, although linguistically meaningful, often failed to align with clinical reality. The descriptive outputs sometimes lacked specificity, and the numeric malignancy scores they produced were inconsistent with ground truth. This highlights the challenge of applying out-of-domain models to high-stakes medical tasks without proper adaptation.

Despite these setbacks, the project offers valuable insights. The modular, agent-based structure proved effective for testing and comparing decision strategies and can be extended further. The system's **interpretability**, even if currently flawed, opens new directions for creating explainable AI tools that can support clinicians—not just with predictions but also with meaningful justifications. To improve future versions, it is essential to **fine-tune the Decision Agents on dermatology-specific datasets**, expand the number of training examples, and explore more **medically aligned prompting strategies**.

# **Conclusion** This project explored the development of a modular, agent-based AI system for classifying dermoscopic images into melanoma and benign categories. By combining transformer-based models—specifically a Vision Transformer (ViT) for triage and CLIP/GIT models for decision support—the system aimed to provide both accurate predictions and interpretable explanations to support clinical decision-making.

# However, the results revealed that the **baseline ViT-only system outperformed the agent-based configuration** in terms of classification accuracy, sensitivity, and specificity. The underperformance of the agent-based system was largely attributed to the **use of general-purpose models (CLIP and GIT)** that were not fine-tuned for the medical domain. Their generated descriptions, while informative linguistically, often lacked clinical relevance, which affected the system’s reliability in diagnosing skin lesions.

# Despite these limitations, the project successfully demonstrates the **feasibility of designing interpretable, agent-driven diagnostic pipelines**. The system architecture provides a flexible foundation for future enhancement, including the incorporation of domain-specific training, improved prompt engineering, and expanded datasets.

# In conclusion, while the current results were less effective than expected, this work provides important lessons on the challenges of applying vision-language models in medical imaging. It reinforces the need for **domain adaptation, clinical validation, and model transparency** in building trustworthy AI systems for healthcare. With further refinement, the agent-based approach holds significant potential to assist clinicians in early melanoma detection and patient education.

# .

# **Acknowledgments**

# I would like to express my sincere gratitude to **Dr. Mutlu Mete**, Professor at the University of North Texas, for his invaluable guidance, mentorship, and constructive feedback throughout the duration of this project. His insights played a crucial role in shaping the direction and depth of this work.

# I also extend my heartfelt thanks to the coordinators of the **UNT-GPREC Health Informatics Internship Program** for providing me with this opportunity to explore the intersection of artificial intelligence and medical diagnostics.

# I am grateful to the faculty and management of **G. Pulla Reddy Engineering College**, Kurnool, for their encouragement and support in pursuing this research.

# 