Interpretable MONET Classification with Concept Graphs and Train-of-Reasoning

Johnathan Zhang

June 10, 2025

Abstract

Modern LLMs have become increasingly relevant in medicine, and dermatology AI models have becoming more accurate in their diagnosis of diseases. However, a common issue in the field is the lack of transparency in how the model found its results. A previous paper introduced MONET[1], a transparent dermatology model that outputs the intermediary steps it took to reach its final answer. We aim to improve the model by introducing a concept graph and a train-of-reasoning output. We propose a pipeline that computes CLIP-based "concept presence" scores for each image over a fixed vocabulary of 56 dermatology concepts, constructs a concept co-occurrence graph and trains a Graph Neural Network (GNN) to refine the scores, uses the refined concept vectors for disease classification, then generates a chain-of-thought explanation via a LLM (T5 or Google Gemini). By training on pubmed dataset derived by the authors of MONET[1] and evaluating on Fitzpatrick17k[2] and DDI[3] datasets, the GNN didn't affect concept-level AUROC much but improved disease-level AUROC by roughly 2%. Legible rationales explaining how detected concepts support and might have led to each disease label. This work demonstrates that the act of combining CLIP, GNNs, and LLMs can potentially improve accuracy and interpretability in dermatology image classification.

1 Introduction

Recently, AI in dermatology have achieved high accuracies in diagnosis but often act as black boxes. Inherently, models learn features by themselves and it is difficult to identify the exact features that they learn off of. This is troublesome since clinicians require transparency and the reasoning that led to a particular diagnosis. Recent work (MONET) leverages CLIP to detect meaningful concepts in dermatology images, then uses the concept scores in a multinomial logistic regression model to classify diseases. However, raw CLIP concept scores can be noisy and ignore inter-concept relationships; some concepts present may not be identified and some absent may wrongly be identified. We address this by building a concept co-occurrence graph over 56 dermatology concepts (48 SkinCon and 7 Derm7pt) and training a simple GNN to refine zero-shot concept scores. This can reduce noise and get rid of some outliers affecting the accuracies. These new refined concept vectors are then used for disease classification with logistic regression as before, then evaluated in terms of its effectiveness with AUROC. Finally, we generate a chain-of-thought explanation using T5 in comparison with Gemini to reason how detected concepts support the predicted disease. This pipeline aims to increase accuracy with interpretability.

2 Methods

2.1 Data and Preprocessing

Training Datasets. The Concept Graph is computed off the following image dataset:

- **Pubmed:** Dermatology image–text pairs (n=105,550) from PubMed articles and medical textbooks collected by original authors of MONET[1].

Evaluation Datasets. We use two publicly available sets with concept annotations:

- **Fitzpatrick17k:** ∼16,000 clinical images, each with 48 binary concept labels and a ground-truth disease label.

- **DDI:** ∼600 clinical images, each with 48 binary concept labels.

Concepts. We define 56 dermatology concepts (48 from SkinCon plus 7 from Derm7pt) to compute concept scores for images, which ranges from shape to color of the condition present in the image.

Concept Scores. We use CLIP ViT-L/14 with weights found by MONET. For each batch of images:

1. Preprocess images sizes to 224 by 224.

2. Encode and normalize images.

3. Tokenize all 56 concepts, encode and normalize them.

4. Compute raw similarity logits by taking the dot product of concepts and images

5. Apply softmax to produce concept scores of shape $\mathbb{R}^{N\times 56}$, with each score $\in$ [0, 1]

2.2 Concept Graph

1. Binarize the concept scores by taking the ones with values over the threshold 0.5

2. Compute co-occurrence: $\mathbf{C} = \text{scores}^\top \times \text{scores}$

3. Build graph G with edges where $\mathbf{C}_{ij} \geq 5000$. We then find the pmi of each edge, where `pmi` $= \log\big(\frac{\mathbf{C}_{ij}/N}{(p_i\,p_j)}\big)$, and $p_k = \frac{\{\text{images containing concept k}\}}{N}$

2.3 GNN

The adjacency matrix from the graph is first normalized, then we use a two-layer GNN to map each image's feature vector x to the GNN-refined scores:

$$m = x + A_{\text{norm}}\,x, \quad h = \text{ReLU}(W_1\,m + b_1), \quad \hat{x} = \sigma(W_2\,h + b_2),$$

where $W_1 \in \mathbb{R}^{256\times 56}$, $W_2 \in \mathbb{R}^{56\times 256}$, and A_{norm} is the normalized adjacency matrix.

Training. The data is split into 80% training and 20% validation splits. The hyperparameters used in training were hidden_dim $= 256$, $\eta = 1 \times 10^{-3}$, $\lambda = 1 \times 10^{-6}$, batch $= 32$, epochs $= 20$. Adam(W_1, b_1, W_2, b_2) optimizer with weight decay λ and MSE loss was used. The final A_{norm} and GNN weights were saved after training. Protoypes of each concept are also saved afterwards for future reference.

2.4 Evaluation

Concept-Level AUROC: annotated fitzpatrick17k and ddi datasets were loaded. We compute the AUROC scores with raw concept vectors generated by MONET and GNN-refined concept vectors. We also average the AUROCs across the categories to see the difference with refinement.

Disease-Level AUROC: We split fitzpatrick17k into 80% and 20% train and test split, then logistic regression models using the either the raw or GNN-refined concept scores associated with the training section of the images. Then, we check the test accuracy using the remaining images with AUROC scores. We also average the AUROCs across the categories to see the difference with refinement.

2.5 Chain-of-Thought

For 5 image samples, we give the predicted disease and GNN-refined concept scores to the model, and ask it to generate a multi-step rationale using the information. Since Fitzpatrick17k was collected in 2022, we generate the rationale using T5. We then compare the T5-generated rationale with a gemini-generated rationale, which is more usable in the real world.

3 Results

3.1 Concept-Level AUROC

- **Zero-shot**: 0.504 ± 0.129
- **GNN-refined**: 0.502 ± 0.137
- **Improvement:** Across several tries, zero-shot and gnn-refined concept-level AUROC remained at the same levels. There's no significant improvement or worsening.

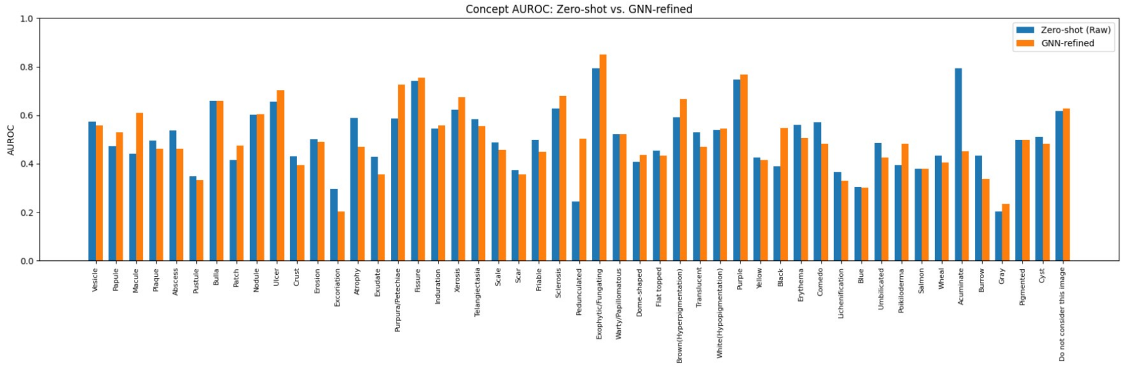


Figure 1: Grouped bar chart of concept-level AUROC: raw (blue) vs. GNN-refined (orange).

3.2 Disease-Level Classification

- **Zero-shot**: 0.828 ± 0.129
- **GNN-refined**: 0.850 ± 0.115
- **Improvement** There is a slight 2.7% improvement but with 7.58% uncertainty.

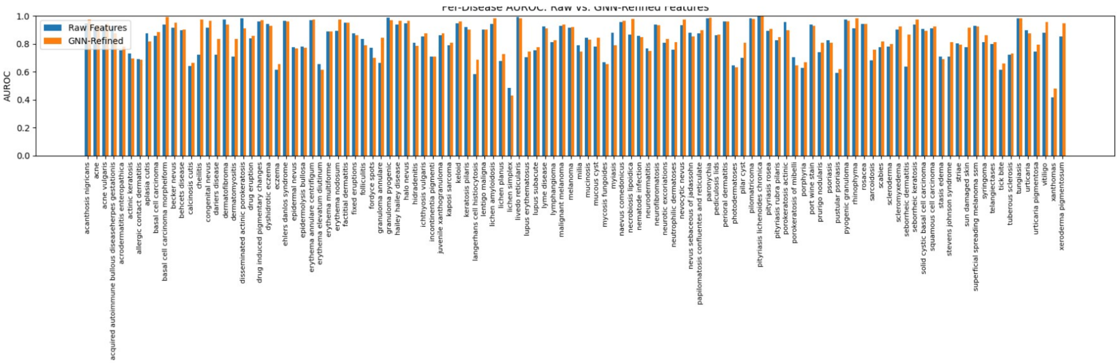
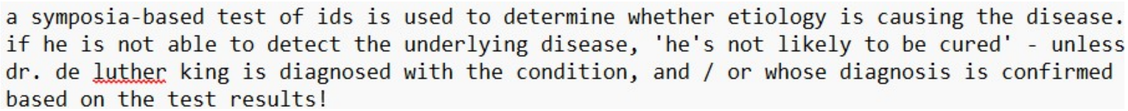


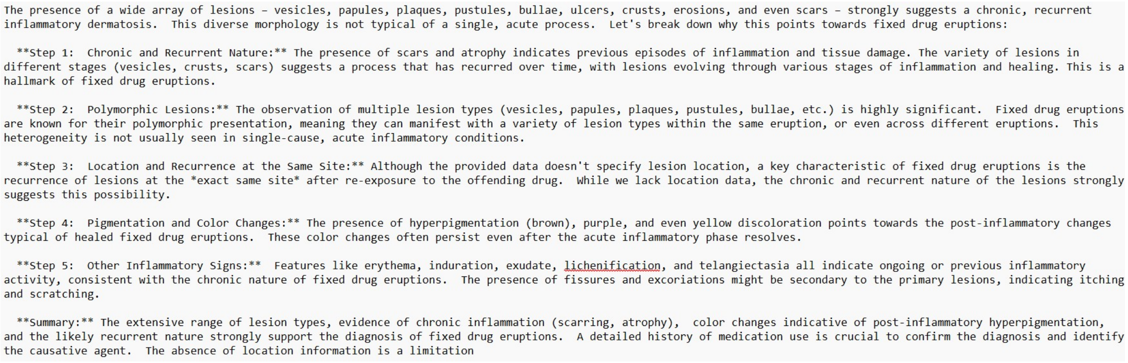
Figure 2: Grouped bar chart of disease-level AUROC: raw (blue) vs. GNN-refined (orange).

3.3 Chain-of-Thought Examples



```
a symposia-based test of ids is used to determine whether etiology is causing the disease.
if he is not able to detect the underlying disease, 'he's not likely to be cured' - unless
dr. de luther king is diagnosed with the condition, and / or whose diagnosis is confirmed
based on the test results!
```

Figure 3: Chain-of-Thought generations with T5-base



```
The presence of a wide array of lesions – vesicles, papules, plaques, pustules, bullae, ulcers, crusts, erosions, and even scars – strongly suggests a chronic, recurrent
inflammatory dermatosis.  This diverse morphology is not typical of a single, acute process.  Let's break down why this points towards fixed drug eruptions:

  **Step 1:  Chronic and Recurrent Nature:** The presence of scars and atrophy indicates previous episodes of inflammation and tissue damage. The variety of lesions in
different stages (vesicles, crusts, scars) suggests a process that has recurred over time, with lesions evolving through various stages of inflammation and healing. This is a
hallmark of fixed drug eruptions.

  **Step 2:  Polymorphic Lesions:** The observation of multiple lesion types (vesicles, papules, plaques, pustules, bullae, etc.) is highly significant.  Fixed drug eruptions
are known for their polymorphic presentation, meaning they can manifest with a variety of lesion types within the same eruption, or even across different eruptions.  This
heterogeneity is not usually seen in single-cause, acute inflammatory conditions.

  **Step 3:  Location and Recurrence at the Same Site:** Although the provided data doesn't specify lesion location, a key characteristic of fixed drug eruptions is the
recurrence of lesions at the *exact same site* after re-exposure to the offending drug.  While we lack location data, the chronic and recurrent nature of the lesions strongly
suggests this possibility.

  **Step 4:  Pigmentation and Color Changes:** The presence of hyperpigmentation (brown), purple, and even yellow discoloration points towards the post-inflammatory changes
typical of healed fixed drug eruptions.  These color changes often persist even after the acute inflammatory phase resolves.

  **Step 5:  Other Inflammatory Signs:**  Features like erythema, induration, exudate, lichenification, and telangiectasia all indicate ongoing or previous inflammatory
activity, consistent with the chronic nature of fixed drug eruptions.  The presence of fissures and excoriations might be secondary to the primary lesions, indicating itching
and scratching.

  **Summary:** The extensive range of lesion types, evidence of chronic inflammation (scarring, atrophy),  color changes indicative of post-inflammatory hyperpigmentation,
and the likely recurrent nature strongly support the diagnosis of fixed drug eruptions.  A detailed history of medication use is crucial to confirm the diagnosis and identify
the causative agent.  The absence of location information is a limitation
```

Figure 4: Chain-of-Thought generations with Gemini

Sample T5-base Generation. This is an extremely bad generation that gives no additional information but has a lot of random text. This might be due to the usage of t5-base, which is a more lightweight model, and doesn't have good reasoning capabilities. The api had been discontinued so we are unable to use a larger version with the time constraint of the project.

Sample Gemini Generation. This generation uses the 1.5 Gemini flash api, which is a newer and much larger model. It evidently had a much better generation and chain-of-reasoning, and can provide valuable insights into why the concepts identified in the image might have led to the diagnosis.

4 Discussion

Although training the GNN was not hyperparameter tuned well due to the short span of time alloted for this project, this project demonstrates that combining a GNN-based refinement step with concept scores can potentially yield higher accuracies in disease identification. Additionally, adding a chain-of-thought explanation using newer LLM models can also help reason why the concept scores may have led to the decisions for the identification, increasing the interpretability. However, we were limited to only 56 concepts and there weren't enough data to capture a lot of different diseases, making the concept graph suffer from the curse of dimensionality.

5 Future Directions

1. **Expand Concept Vocabulary.** More than 56 concepts are needed for better diagnosis accuracy.

2. **Hyperparameter tuning.** Better hyperparameter tuning can be done on the GNN to increase accuracy.

3. **Outside validation.** Have a domain expert annotate a subset of clinical images to evaluate zero-shot concept detection and disease classification on unseen data.

4. **Data.** With more annotated data possibly released in the future, the accuracies of the model can continuously improve.

6 Citations

[1] Kim, C., Gadgil, S.U., DeGrave, A.J. et al. Transparent medical image AI via an image–text foundation model grounded in medical literature. Nat Med 30, 1154–1165 (2024). https://doi.org/10.1038/s41591-024-02887-x

[2] Groh, M., Harris, C., Soenksen, L., et al. (2021). Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1820-1828).

[3] Roxana Daneshjou et al. ,Disparities in dermatology AI performance on a diverse, curated clinical image set.Sci. Adv.8,eabq6147(2022). https://doi.org/10.1126/sciadv.abq6147