Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2025) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

Eighth International Conference Futuristic Trends in Networks and Computing Technologies (FTNCT08) with Organizing university venue: KIET Group of Institutions(KIET), Ghaziabad, India

An Open and Reproducible Multimodal AI Framework for Skin Disease Diagnosis Using Public Vision and Biomedical Language Models

Chola Chetan Chukkala[a], David Shibu[a], M.G. Thushara[b,*], Nikhilesh Krishna Chukkala[a]

[a]*Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri, India*
[b]*Department of Computer Science and Applications, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri, India*

Abstract

The accurate diagnosis of dermatological conditions remains a persistent challenge owing to the high inter-class visual similarity and intra-class variability among skin lesions. Although deep learning models have demonstrated considerable progress, unimodal systems that depend exclusively on image data often fail to capture the contextual and semantic cues embedded in medical narratives. This study presents MAGE-V-Pro (Multimodal Attention-Guided Encoder for Vision), a novel multimodal framework designed to enhance dermatological disease classification. The proposed model adopts a two-stage training strategy: initially, a vision-domain adaptation phase where a DINOv2 Vision Transformer is fine-tuned on dermatology-specific image data, followed by a multimodal fine-tuning phase that integrates vision features with biomedical text embeddings derived from PubMedBERT through a Feature-wise Linear Modulation (FiLM) layer. To achieve such a computational efficiency, we applied Low-Rank Adaptation (LoRA) to the text encoder. Also,this framework is evaluated on a balanced dataset having seven dermatological conditions (images which are publicly available) along with patient-style text descriptions (in patient point of view) generated for reproducibility. MAGE-V-Pro has attained a test accuracy of 95.71%, promising a huge improvement over the 71.43% achieved by its vision-only model. These findings show that combining medical images along with text description provides an improved diagnostic accuracy and enhanced interpretability through FiLM-based textual attention and Grad-CAM visual analyses.

Keywords: Multimodal Learning; Dermatological Disease Classification; Vision Transformer; PubMedBERT; FiLM Fusion; Deep Learning; MAGE-V-Pro

* Corresponding author: M.G. Thushara. Email: thusharamg@am.amrita.edu

1. Introduction

Computer-aided skin disease classification has been prominent in computational medicine and promises scalable and reproducible systems. With advances in deep learning and crafting good vision transformers, processing of medical imagery has gone through a drastic transformation [5]. But dermatology happens to be a very challenging field with its very nature of visual complexity being characterized by *high inter-class similarity*—where lesions that look similar can actually be caused by very different diseases and *high intra-class variation*, meaning the same disease can look different from person to person.

Entirely based on vision data is practically limiting since it ignores rich contextual and semantic data that clinicians rely on in the practice of clinical diagnosis in real life. Natural language descriptions of how a lesion looks along with symptom details, often provide key clues for identifying cases that look similar on the surface [19]. What is truly necessary is the building of **multimodal models** which are capable of processing both visual and textual data at once in an uncomplicated manner with the hopes of forming more complete and explainable analyses [14]

To address this requirement, we introduce **MAGE-V-Pro (Multimodal Attention-Guided Encoder for Vision)**, a new multimodal system aimed at improving dermatological disease classification. This system undergoes a two-phase training process: first, a strong vision model is introducedin the domain of dermatology, and then it is combined with a biomedical text encoder in order to attain multimodal learning. By this fusion process, *MAGE-V-Pro* conditions visual reasoning accordingto textual context, thus replicating the diagnostic process of the expert.

The key findings of this work are the following:

1. **Two-Stage Training Approach:** A vision-only **DINOv2** model [16] is first fine-tuned using a large image dataset of Dermnet (only the seven classes excluding Stage - 2 images within the train directory) with class labels and not textual data. We choose Dermnet because it is publically available and widely accessible. In the second stage, the complete multimodal **MAGE-V-Pro** system is fine-tuned on paired image–text data (image + text). This stage by stage training process effectively addresses the imbalance between abundant image data instances and limited textual data.
2. **Text-Conditioned Visual Attention:** A **Feature-wise Linear Modulation (FiLM)** layer helps for *text-conditioned visual attention*, where embeddings from **PubMedBERT** adjust the visual features extracted from **DINOv2**, resulting in a **text-guided vision transformer**.
3. **Confidence-Based RAG Mechanism:** For improved reliability, the system includes a **Confidence-Based Retrieval-Augmented Generation (RAG) Fallback** for low-confidence predictions. When model confidence falls below a predefined threshold, relevant clinical information is retrieved from a filtered and reviewed **knowledge base**[15], providing a secondary reference to assist in decision-making.
4. **Replicable Patient Descriptions Using the LLM for Controllable Dataset Creation:** To mitigate the shortage of multimodal datasets, patient-style textual descriptions are generated using a large-scale language model (**Gemini**). This process produces a **reproducible** and **high-quality image–text dataset**, facilitating **transparent** and **equitable benchmarking and evaluation**.To ensure that bias is minimised, every description was reviewed thoroughly, and the prompt given to gemini clearly ensured the creation of non-leaky text descriptions.

The remaining part of the paper lays out the following structure: Section 3 presents the methodology used, Section 4 lays out the plan and findings of the experiments, Section 5 analyzes the findings and potential applications, and Section 6 summarizes the work with concluding statements and future work directions.

2. Related Work

The current work intersects three large fields: **deep learning applications in dermatology**, **multimodal fusion techniques**, and **large-scale pre-trained base models**. We briefly describe each field and point out the research gaps which motivate the current work.

2.1. Deep Learning in Dermatological Education

The use of **Artificial Intelligence (AI)** in dermatological diagnosis has advanced substantially over the past decade. Early breakthroughs were achieved through the use of **Convolutional Neural Networks (CNNs)** for automatic classification of skin lesions. A seminal study by *Esteva et al.* [5] demonstrated that CNNs can reach dermatologist-level accuracy in skin cancer detection, sparking a wave of AI-driven research in dermatology ranging from external skin problems to various internal issues in the gastrointestinal tract. [11]. Subsequent studies extended these approaches to varied architectures and applications, including **multimodal human–AI diagnostic systems** [21] and treatment planning support tools[18].

The recent migration from convolutional models towards **Vision Transformers (ViTs)**[4] also transformed medical image analysis, owing to the fact that ViTs are capable of encoding the entire spatial structure in an image at the global level. Owing to such a transformation, the emergence of **large-scale self-supervised pretrained models** such as **DINOv2**[16] allows learning generalizable and rich visual features from large-scale unannotated datasets. Such models can also serve as a helpful initialization point in domain adaptation in specialized medical domains such as dermatology, and are the cornerstone for the implementation process of the **vision-domain adaptation process** in our work.

2.2. Multimodal Fusion in Medical Imaging

Although image-based models are still cornerstones in computer-aided diagnosis, practice-level clinical reasoning is fundamentally **multimodal**. As a part to this fact, recent work has therefore sought the fusion of auxiliary patient information with image information in the aim of improving diagnostic performance. Initial dermal examples paired visual characteristics with **tabular metadata**—e.g., patient age, sex, or lesion site—yielding quantifiable advances over the state-of-the-art image-only systems [13].

But the fusion of **unstructured textual information** becomes even more challenging. The older fusion techniques, i.e., **late fusion** (prediction averaging), and **early fusion** (concatenation of features), can rarely recover the semantic dependencies among modalities.

Distinct from these, our approach leverages **Feature-wise Linear Modulation (FiLM)**[17], a conditioning technique originally proposed for visual reasoning. In FiLM, textual embeddings modulate visual feature activations adaptively, guiding attention based on linguistic context. The application of FiLM as a **text-driven attention mechanism** for dermatological image classification which remains relatively underexplored.

2.3. Biomedical Language Models and Parameter-Efficient Fine-Tuning

The textual component of our system benefits from recent advancements in **Natural Language Processing (NLP)**. The introduction of the **Transformer architecture**[22] and the emergence of **large pre-trained models** such as **BERT**[3] have revolutionized language understanding. Subsequent domain-specific pre-training significantly improved performance in biomedical tasks, leading to models such as **BioBERT**[12] and **PubMedBERT**[7], both trained on large-scale biomedical text corpora.

In our system, the use of **PubMedBERT** allows the language encoder to access embedded medical vocabulary and contextual understanding. Fine-tuning such large models, however, results in computational and memory costs. This was countered by the introduction of **Parameter-Efficient Fine-Tuning (PEFT)** techniques, which enable model adaption with fewer trainable parameters.. Early PEFT methods, such as **Adapters**[9], were later advanced into more efficient methods such as **Low-Rank Adaptation (LoRA)**[10]. LoRA introduces compact, trainable low-rank matrices in the model, enabling efficient fine-tuning with no tradeoff in representational capability. In our own case, therefore, this makes **LoRA** extremely fit for adapting PubMedBERT in our **multimodal dermatological framework**. While the latest models like Med PaLM 2[20] and others have multimodal capabilities, this research aims to work on Open Source Models.

3. Methodology

Our proposed research focuses on the **MAGE-V-Pro** model, a multimodal architecture created especially to diagnose dermatological diseases. In the next section, we describe the datasets and preprocessing, the model architecture, and the two-stage training procedure used to fine-tune the precision of diagnosis.

3.1. Dataset and Preprocessing

To provide maximal reproducibility and transparency, only public datasets have been considered in this work. The initial source of images was the **DermNet** repository, which happens to be an extremely well-curated collection of dermatological images that cover an incredible number of skin ailments. The dataset was divided into two sets depending on whether corresponding textual data was available or not:

- **Vision-only set (Stage 1):** The set consisted of 5,605 textually labeled images without textual annotations and was utilized to train the vision encoder on dermatology-specific visual representations.
- **Multimodal set (Stage 2):** With 2,030 image–text pairs (exclusively of the test samples), this set served to fine-tune the complete multimodal **MAGE-V-Pro** model.

A strictly set-apart **test set** containing 70 image–text pairs was reserved for final assessment to ensure unbiased estimation of performance. This test set's text component contained patient-type clinical notes generated by a large language model (Gemini) where the prompt is taken care to keep the text non-leaky and the descriptions were gone through in detail, thereby approximating in-real-life diagnosis notes. Seven representative skin ailments, including acne, psoriasis, eczema, sexually transmitted infections (STDs), fungal infections, basal cell carcinoma (BCC), and seborrheic keratosis, were targeted in the study.

To encourage generalization, routine data augmentation processes at training were employed, i.e., random cropping, horizontal and vertical flips, rotation, and color jitter. Images were down-sized to 224×224 pixels and normalized to the vision encoder's input specifications.

3.2. Model Architecture

The **MAGE-V-Pro** architecture (see Figs. 1 and 2) brings together three basic modules: a text encoder, a vision encoder, and a multimodal fusion module that enables joint learning between both modalities.

Vision Encoder: Vision backbone consists of a self-supervised **DINOv2** model (`dinov2_vits14`) that is capable of learning powerful, semantically relevant features in an unsupervised manner. The output of the encoder is a 384-D feature embedding per image, capturing both local and global lesion characteristics.

Text Encoder: With text modality and **PubMedBERT**, this Transformer was pretrained on vast biomedical data. With an objective to restrict computational expense and still provide adaptability, **Low-Rank Adaptation (LoRA)**is employed to add sparse, learnable parameters with the remaining model weights frozen.

FiLM Fusion Module: Combining the two modalities employs a **Feature-wise Linear Modulation (FiLM)** layer. This module enables text-conditioned feature modulation by generating two parameter vectors—an optional scale factor (γ) and an optional bias term (β)—from the text embeddings. Simple concatenation treats all modalities equally whereas FiLM allows to dynamically dynamically adjusts the visual feature weights based on the given text context. The two vectors apply the modulation to the visual representations in the following manner:

$$F_{\text{fused}} = \gamma \odot V_{\text{features}} + \beta \tag{1}$$

where $\odot$ symbolizes element-wise product. After that, the resultant representation having been combined goes to a Multi-Layer Perceptron (MLP) classifier to give the final diagnostic output.

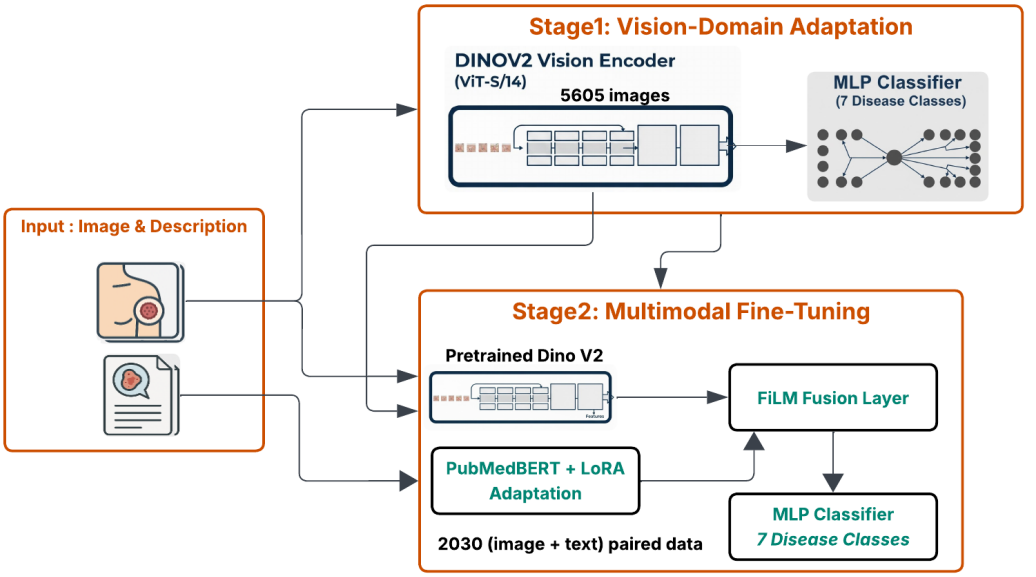


Fig. 1: Architecture of MAGE-V-Pro showing DINOv2 vision adaptation in Stage-1 and multimodal image–text fusion with FiLM and LoRA-adapted PubMedBERT Stage-2.

3.2.1. Retrieval-Augmented Generation (RAG)

In order to increase reliability in ambiguous cases, **MAGE-V-Pro** also has a feature of refinement based on *Retrieval-Augmented Generation (RAG)*. Whenever the model's confidence in an initial hypothesis drops below a set threshold of 0.6 (set in our model), it interacts with a codified dermatological knowledge base. This necessitates the system to fetch and display differential questions to the user. Answers to these questions help to make distinctions between clinically similar conditions and hence allow the system to refine its prediction. This early version serves only as an initial step, with plans to create more advanced interaction reasoning in future works.

3.3. Two-Stage Training Strategy

Training on models was conducted in two subsequent steps to achieve full potential in each of unimodal adaptation and multimodal combination.

Stage 1 – Vision-Domain Adaptation: In the first stage, the DINOv2 encoder was trained beforehand on the vision-only dataset to transfer its pre-trained representation to dermatology-related visual patterns. Weighted cross-entropy loss was used to minimize class imbalance. At this stage, a maximum validation accuracy of 72.88% was achieved, which served to establish a good vision-only baseline.

Stage 2 – Multimodal Fine-Tuning: During Stage 2, the complete **MAGE-V-Pro** model was fine-tuned with the fine-tuned vision encoder of Stage 1. Simultaneous optimization of the multimodal dataset was carried out under differential learning rates—smaller on the vision encoder and higher on the parameters of the LoRA and MLP—to make trade-offs between stability and flexibility. Stage 2's highest validation accuracy was 98.52%, and it clearly showed the advantage of multimodal learning over unimodal baselines.[8]

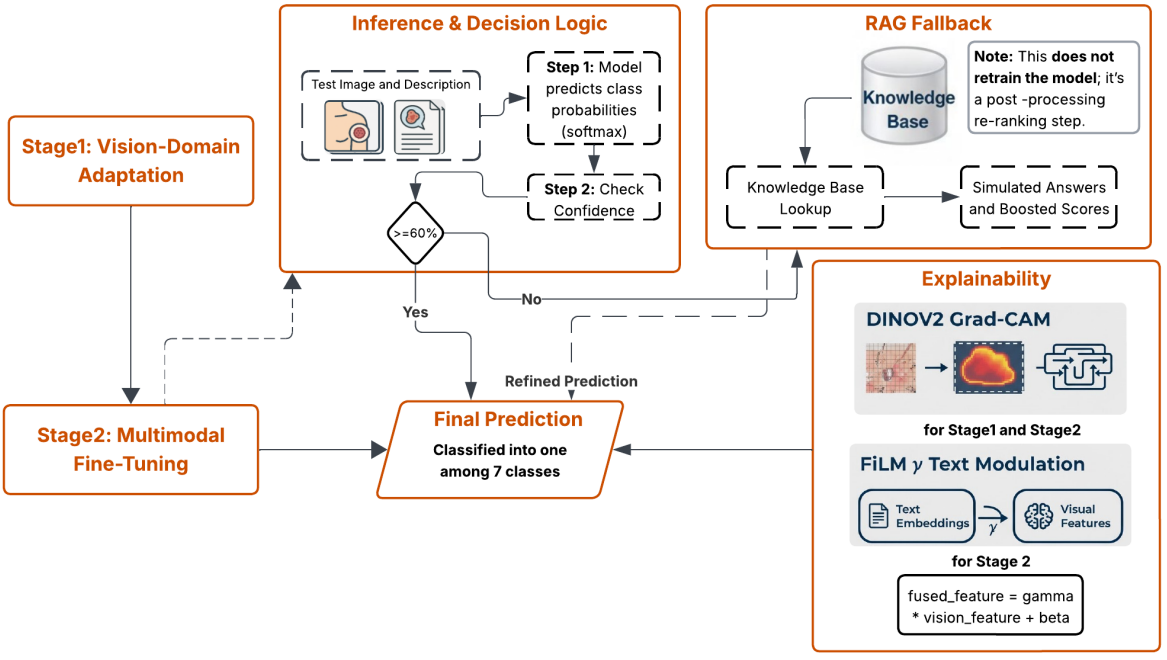


Fig. 2: Continuation of the architecture illustration of MAGE-V-Pro, showing extended module flow and detailed processing steps, including Inference logic and knowledge based RAG model.

4. Experiments and Results

We present the experimental setting, the performance criteria, and the empirical analysis of the proposed **MAGE-V-Pro** framework in this section. To demonstrate the performance advantage of the proposed multimodal approach over a strong baseline of the single-modal method, the experiments were conducted.[2][6]

4.1. Implementation Details

Experiments were performed in the **PyTorch** environment. Visual backbone (`dinov2_vits14`) was initialized from the PyTorch Hub, and the text encoder (`PubMedBERT`) was imported with the **Hugging Face Transformers** library along with the **PEFT** module for LoRA-based fine-tuning. Training was done only with one NVIDIA Tesla T4 GPU.

Training followed the two-stage protocol described in Section 3. In **Stage 1** (vision-domain adaptation), the DINOv2 backbone was trained for 20 epochs with batch size 32 and with learning rate 1×10^{-4}. In **Stage 2** (multimodal fine-tuning), the text and vision streams were trained together for 30 epochs with batch size 16. Differential learning rates were employed: 1×10^{-6} for the frozen DINOv2 encoder and 5×10^{-5} for the LoRA params and the MLP classifier. We used the **AdamW** optimizer at all times and early stopping (patience 4 and 5 epochs in the two stages, respectively) for preventing overfitting.

4.2. Evaluation Metrics

A held-out test set comprising 70 image–text pairs was also used to assess the model's performance. Per-class precision, recall, F1-score, and overall accuracy were among the standard classification measures that were calculated. When taken separately, these indicators allow for a thorough assessment of both the overall and category-level performance across the seven dermatological classes.

4.3. Performance Analysis

Table 1: Performance of the vision-only baseline on the test set.

Class	Precision	Recall	F1-Score	Support
acne	0.9000	0.9000	0.9000	10
psoriasis	0.6154	0.8000	0.6957	10
eczema	0.7500	0.6000	0.6667	10
stds	1.0000	0.4000	0.5714	10
fungal	0.8889	0.8000	0.8421	10
bcc	0.4286	0.6000	0.5000	10
seborrheic_keratosis	0.7500	0.9000	0.8182	10
Accuracy			**0.7143**	70
Weighted Avg	0.7618	0.7143	0.7134	70

The baseline vision-only reached an average accuracy of **71.43%**. From Table 1, the model excelled at largely differentiated classes like `acne` (F1 = 0.90), but failed at morphologically overlapping conditions like `bcc` and `stds`. We can see from the confusion matrix in Fig. 3 that such misclassifications mainly happened among morphologically related.

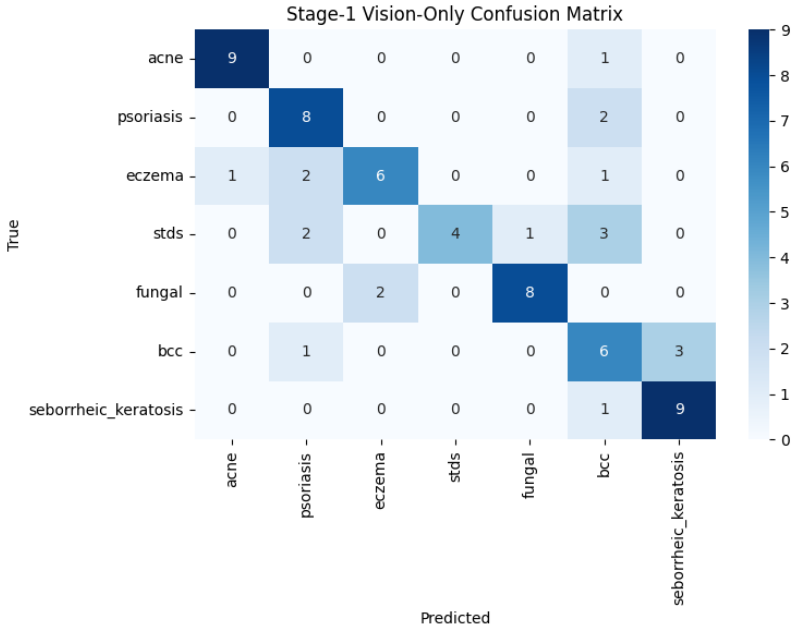


Fig. 3: Confusion matrix for the vision-only baseline model. Notable confusion is observed between `bcc` and `seborrheic_keratosis`.

The proposed **MAGE-V-Pro** system achieved a test set accuracy of **95.71%** with absolute gain of 24.28 percentage points over the baseline unimodal system. Tabulated in Table 2 are the system precision and recalls which were very close to perfect in nearly all classes. From the confusion matrix in Fig. 4 it can be observed that the multimodal model demonstrated increased discriminative capability with a considerable reduction in cross-class confusion. Notably, the F1 scores has a significant increase almost near to 1 which is most important in medical domains as compared to Table 1 and Table 2 results.

Table 2: Performance of the MAGE-V-Pro multimodal model on the test set.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
acne	1.0000	1.0000	1.0000	10
psoriasis	0.9091	1.0000	0.9524	10
eczema	1.0000	0.8000	0.8889	10
stds	1.0000	1.0000	1.0000	10
fungal	0.9000	0.9000	0.9000	10
bcc	0.9091	1.0000	0.9524	10
seborrheic_keratosis	1.0000	1.0000	1.0000	10
Accuracy			**0.9571**	70
Weighted Avg	0.9597	0.9571	0.9562	70

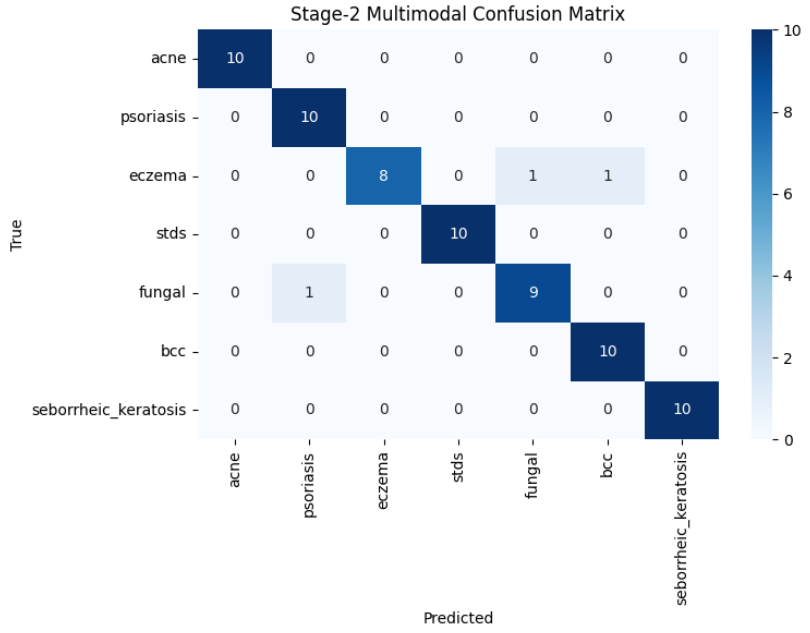
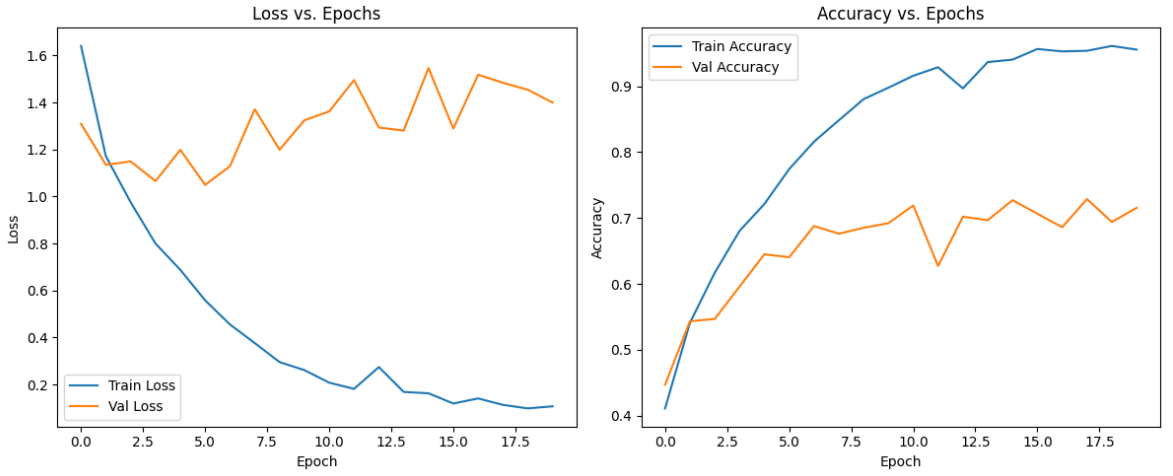


Fig. 4: The confusion matrix for the multimodal MAGE-V-Pro model with high class discrimination compared with the baseline vision-only model.

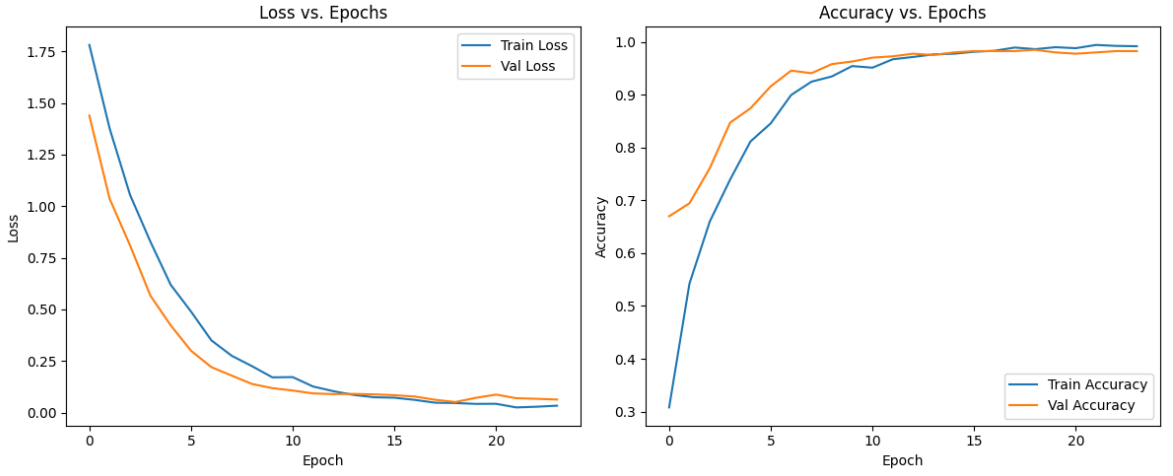
4.4. Training Dynamics and Qualitative Analysis

The learning procedures for the two stages of training are depicted in Fig. 5. The unimodal (vision-only) model converges early with a plateau in the validation at roughly 70%, while the multimodal MAGE-V-Pro converges more smoothly and ends up above the 98% level in the validation accuracy. Such smoothness can be accounted for with the auxiliary contextual information incorporated in the textual modality.

Fig. 6 illustrates qualitative attention visualizations derived from the FiLM γ weights for representative samples from each of the seven disease categories. These visualizations act as modality-aware attention indicators, highlighting the visual features most influenced by corresponding textual descriptions. Different activation patterns across classes indicate that **MAGE-V-Pro** learns to use text information while analyzing images. This makes the model easier to understand and helps it think more like a doctor



(a) Training and validation curves for the vision-only model.



(b) Training and validation curves for the multimodal MAGE-V-Pro model.

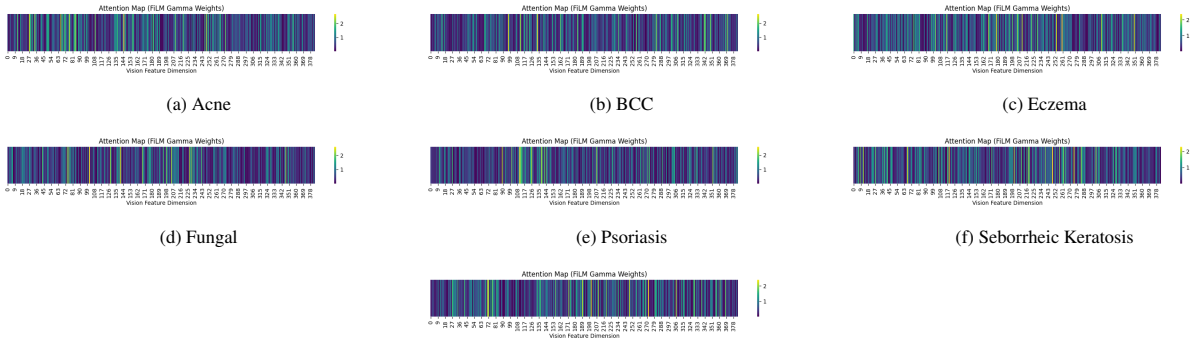
Fig. 5: Comparison between training procedures from the unimodal and the multimodal periods. The multimodal model converges more rapidly and consistently with the augmenting textual feedback.

4.5. Saliency Mapping with Grad-CAM

To also examine and visually verify the models' choice-making processes, we utilized Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM generates a heatmap which indicates the precise pixels in an input image that are most informative for a specific prediction. In examining the heatmaps from the vision-only baseline and the multimodal **MAGE-V-Pro** models, we can immediately see how textual context informs the model's visual attention.

The outcomes, shown in Fig. 7, reveal the possible advantages of multimodal fusion. In multiple classes, i.e., `acne`, `bcc`, `fungal`, and `seborrheic_keratosis`, the models accurately localize the lesion and yield localized heatmaps at the exact position. Hence, the superiority of **MAGE-V-Pro** becomes evident in visually complex cases.

One such example includes the `stds` sample, in which the vision-only model incorrectly identifies the inflammation as `psoriasis`. In contrast, **MAGE-V-Pro**, with the aid of the supporting text, accurately identifies the condition and produces a more accurate heatmap. What this shows is the capacity of the model to use textual descriptions in order to overcome visual ambiguity and remove all misdiagnosis (considerably). What the analysis also shows are ongoing challenges; for example, the `eczema` sample was misclassified in the models, which shows that some difficult textures



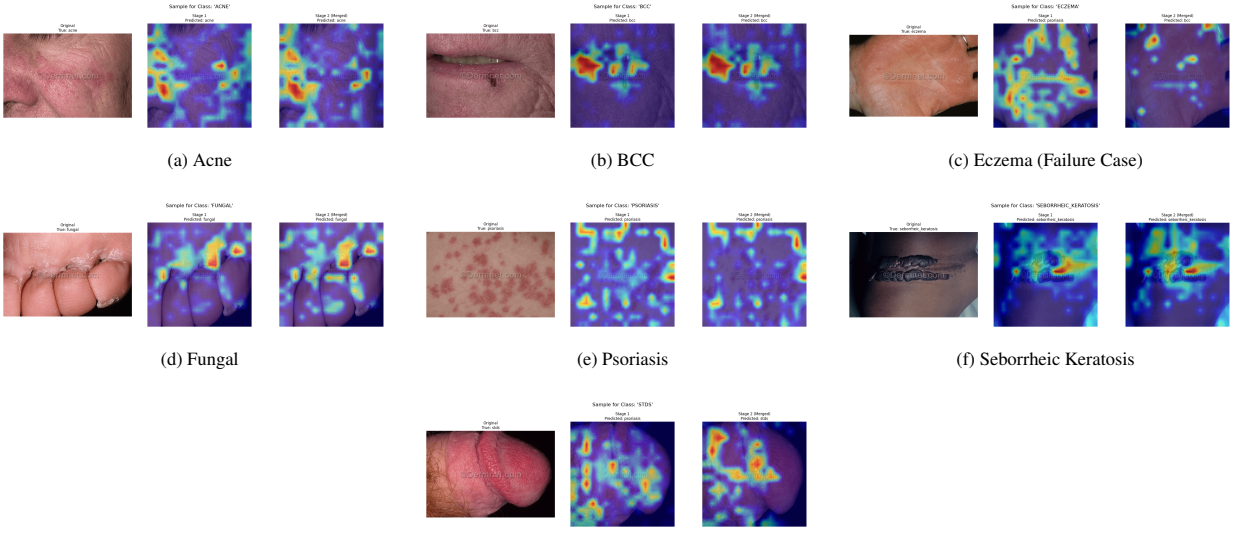
(a) Acne (b) BCC (c) Eczema

(d) Fungal (e) Psoriasis (f) Seborrheic Keratosis

(g) STDs

Fig. 6: This figure shows FiLM γ weights for sample images from all seven classes and these γ values work like attention signals, adjusting all the image features (384 in total) based on the text. The different patterns across classes show that MAGE-V-Pro focuses on the most useful visual details for each disease in a way that makes medical sense.

are still hard to differentiate even with the use of multimodal context. Generally, the Grad-CAM analysis verifies the fact that the **MAGE-V-Pro** actually trains to base its visual reasoning in a more text (patient here) applicable fashion so that improved accuracy and interpretability are achieved.



(a) Acne (b) BCC (c) Eczema (Failure Case)

(d) Fungal (e) Psoriasis (f) Seborrheic Keratosis

(g) STDs (Correction Case)

Fig. 7: Grad-CAM visualizations for one sample in each of the seven test classes. Each pair displays the original image, the heatmap from the Stage-1 vision-only model, and the heatmap from the Stage-2 MAGE-V-Pro model. The STDs case stands out in highly revealing MAGE-V-Pro correcting the baseline's misclassification.

5. Discussion

The findings from our experiments in Section 4 presents the experimental results illustrating the performance of the proposed **MAGE-V-Pro** framework in supporting multimodal dermatological diagnosis. A clear improvement in accuracy from 71.43% to 95.71% illuminates that text-guided feature conditioning decidedly augments the ability of vision transformers to discriminate in advanced medical imaging tasks. This section presents an interpretive analysis

of these results, discusses the qualitative model behavior in more detail, and describes the main shortcomings and future research prospects.

5.1. Effect of Multimodal Fusion on Diagnostic Accuracy

By far the most noteworthy result of this research is the very notable performance difference between the unimodal and multimodal configurations. A look at the confusion matrix of the vision-only setting (Figure 3) finds that visually comparable states—like basal cell carcinoma (`bcc`) and seborrheic keratosis(`seborrheic_keratosis`)—were frequently misclassified, highlighting the built-in arbitrariness of using morphologic features alone. Similar overlaps were also found between `stds` and `psoriasis`, also emphasizing the arbitrariness of distinguishing between diseases with overlapping visual characteristics.

When textual data was added, these uncertainties were reduced at a high way. In Figure 4, we observe that the MAGE-V-Pro model had almost perfect recall on all classes and practically no `bcc` and `stds` misclassifications. This improvement reveals the usefulness of textual context in guiding visual interpretation. Text descriptions with significant linguistic cues—such as ”pearly,” ”non-healing,” or ”crusted”—help guide the model to the correct diagnosis even with high visual similarity. FiLM-based conditioning makes this process possible by adaptively tuning visual feature representations in accordance with textual data (the context here) and thereby approximating the combined reasoning process undertaken by clinicians.

5.2. Qualitative Analysis of FiLM Attention

To more explicitly illustrate the fusion dynamics, the γ weights of the FiLM layer were also explored (Figure 6). They function as a form of feature-level attention, gating how textual context influences the activation of visual features. Their distinctive patterns of activation that register across disease classes suggest class-specific modulatory behaviors learned by the model. Textual inputs that specify ”silvery scales” in `psoriasis`, e.g., produce feature activations that vary compared to those that specify ”weeping blisters” in `eczema`. Although the FiLM attention model does not provide us with the same level of fine-grained spatial localisation that Grad-CAM achieves, it still derives rich semantic alignment between textual and visual representations to permit feature-level interpretable reasoning. We will include an ablation comparison study between FiLM and simple concatenation in the extended work. The selected modulation confirms that the model focuses on the most important visual features and effectively removes the noise.

5.3. Limitations and Future Work

While the results have potential, a number of weaknesses ought to be acknowledged to guide future studies and clinical translation:

1. **Dataset Diversity and Count:** A small subset of public dataset (Dermnet) is used in this research to make the results reproducible within limited clinical and image diversity. Future evaluations will be focusing on utilize larger, heterogeneous datasets that vary in skin pigmentation, type of imaging, and clinical scenario with more chances of working in the domain including a focus of lesser count data using zero-shot learning and few-shot learning[1].
2. **Text Modality Source:** Test set text modality was created with an LLM (Gemini), which may carry stylistic biases that vary from clinical narratives in everyday life. Future research ought to confirm model generalizability on real patient notes or Electronic Health Record (EHR) data to make it more robust in real clinical settings. The textual generated data will be made open soon.
3. **Confidence-Based RAG Framework:** Although the architecture incorporates a Confidence-Based Retrieval-Augmented Generation (RAG) fallback system, its empirical merit was quantitatively not verified during this phase. Merging and benchmarking this module is an enriching future direction toward a more interpretable and reliable diagnosis system, in particular under out-of-distribution or low-confidence conditions along with .

Other than these areas, future studies shall generalize fairness and mitigation of biases among population subpopulations and extend the model to permit explainable multimodal reasoning in additional clinical applications of medical imaging. In the current paper, we are proposing a framework which can be further modelled for a real clinical environment. Overall, such directions shall make the proposed **MAGE-V-Pro** model even more clinically practical, safe, and interpretable.

6. Conclusion

In this study, we present **MAGE-V-Pro**, a multimodal system constructed to address the inadequacies inherent with vision-based approaches to dermatological image classification. In adding both text and image modalities within a systematic two-stage training procedure, our proposed system demonstrates the capability of vision transformer-based domains to significantly benefit from biomedical understanding of text, leading to clinically relevant performance gains.

In the subsequent step, contextual embeddings learned from the PubMedBERT language model were combined with fine-graded visual features in a Feature-wise Linear Modulation (FiLM) layer to permit feature modulating based on text-based inputs. This architecture enables the model's ability to emulate clinical reasoning by adaptively modulating its visual focus based on patient-individual textual rationales.

Tests that used publicly available datasets confirmed the capacity of the framework proposed. Our model had a test accuracy of **95.71%** and dominated the strong vision-only baseline (71.43%) by **24.28 percentage points**. This huge gain at the base highlights the importance of multimodal fusion in dealing with the visually complicated understandings.

Beyond performance, MAGE-V-Pro focuses on reproducibility and interpretability which is very keen in deep learning papers. With publicly available data and systematically generated patient descriptions, we establish an open and extendable benchmark to future multimodal medical AI studies. In addition, the FiLM-based attention scheme presents interpretable in-between results and bridges algorithmic decisions with clinical thinking. We encourage in focusing more on how to make correct decision when the text description is incomplete and incorrect.

Future research will extend this foundation by improving the Confidence-Based Retrieval-Augmented Generation (RAG) module, with the goal of enhancing reliability under conditions of uncertainty or all unexpected scenarios. In summary, MAGE-V-Pro pushes an important step toward clinically reliable, interpretable, and replicable multimodal artificial intelligence systems for dermatological diagnosis and a framework that can be similar in other fields of medical imaging.

References

[1] Anjali, T., Abhishek, S., Varnika, V.N., Abhinav, V., 2025. Real-time portable diagnostics for seborrheic dermatitis via hierarchical few-shot learning. IEEE MultiMedia , 1–16doi:10.1109/MMUL.2025.3623700.

[2] Baltruaitis, T., Ahuja, C., Morency, L.P., 2019. Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 423–443. doi:10.1109/TPAMI.2018.2798607.

[3] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding, in: Burstein, J., Doran, C., Solorio, T. (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota. pp. 4171–4186. URL: https://aclanthology.org/N19-1423/, doi:10.18653/v1/N19-1423.

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=YicbFdNTTy.

[5] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118. URL: https://doi.org/10.1038/nature21056, doi:10.1038/nature21056.

[6] Gao, J., Li, P., Chen, Z., Zhang, J., 2020. A survey on deep learning for multimodal data fusion. Neural Computation 32, 829–864. URL: https://doi.org/10.1162/neco_a_01273, doi:10.1162/neco_a_01273, arXiv:https://direct.mit.edu/neco/article-pdf/32/5/829/1865303/neco$_{a0}$1273.*pdf*.

[7] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H., 2021. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans. Comput. Healthcare 3. URL: https://doi.org/10.1145/3458754, doi:10.1145/3458754.

[8] Hari Narayanan, A.G., Amar Pratap Singh, J., 2021. Skin disease ensemble classification using transfer learning and voting classifier. International Journal of Engineering Trends and Technology 69, 287–293. URL: https://doi.org/10.14445/22315381/IJETT-V69I12P234, doi:10.14445/22315381/IJETT-V69I12P234.

[9] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S., 2019. Parameter-efficient transfer learning for NLP, in: Chaudhuri, K., Salakhutdinov, R. (Eds.), Proceedings of the 36th International Conference on Machine Learning, PMLR. pp. 2790–2799. URL: https://proceedings.mlr.press/v97/houlsby19a.html.

[10] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2021. Lora: Low-rank adaptation of large language models. URL: https://arxiv.org/abs/2106.09685, arXiv:2106.09685.

[11] Kumar, S.Y., J M, A.G., Vimal, T., Abhishek, S., T, A., 2025. Beyond x-ray: Deep learning solutions for gastrointestinal bleeding. Procedia Computer Science 259, 1306–1315. URL: https://www.sciencedirect.com/science/article/pii/S187705092501186X, doi:https://doi.org/10.1016/j.procs.2025.04.085. sixth International Conference on Futuristic Trends in Networks and Computing Technologies (FTNCT06), held in Uttarakhand, India.

[12] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J., 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36, 1234–1240. URL: https://doi.org/10.1093/bioinformatics/btz682, doi:10.1093/bioinformatics/btz682, arXiv:https://academic.oup.com/bioinformatics/article-pdf/36/4/1234/48983216/bioinformatics$_3$6$_4$1234.*pdf*.

[13] Liu, Y., Jain, A., Eng, C., Way, D.H., Lee, K., Bui, P., Kanada, K., de Oliveira Marinho, G., Gallegos, J., Gabriele, S., Gupta, V., Singh, N., Natarajan, V., Hofmann-Wellenhof, R., Corrado, G.S., Peng, L.H., Webster, D.R., Ai, D., Huang, S.J., Liu, Y., Dunn, R.C., Coz, D., 2020. A deep learning system for differential diagnosis of skin diseases. Nature Medicine 26, 900–908. URL: https://doi.org/10.1038/s41591-020-0842-3, doi:10.1038/s41591-020-0842-3.

[14] Mohan, J., Sivasubramanian, A., V., S., Ravi, V., 2025. Enhancing skin disease classification leveraging transformer-based deep learning architectures and explainable ai. Computers in Biology and Medicine 190, 110007. URL: https://www.sciencedirect.com/science/article/pii/S0010482525003580, doi:https://doi.org/10.1016/j.compbiomed.2025.110007.

[15] N, P., T, R., Thushara, M., Krishna, K.A., V, P., 2025. Retrieval-augmented generation for multiple-choice questions and answers generation. Procedia Computer Science 259, 504–511. URL: https://www.sciencedirect.com/science/article/pii/S1877050925010968, doi:https://doi.org/10.1016/j.procs.2025.03.352. sixth International Conference on Futuristic Trends in Networks and Computing Technologies (FTNCT06), held in Uttarakhand, India.

[16] Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P., 2024. DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research URL: https://openreview.net/forum?id=a68SUt6zFt. featured Certification.

[17] Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A., 2018. Film: Visual reasoning with a general conditioning layer. Proceedings of the AAAI Conference on Artificial Intelligence 32. URL: https://ojs.aaai.org/index.php/AAAI/article/view/11671, doi:10.1609/aaai.v32i1.11671.

[18] Rauniyar, A., Reddy, B.M., S, R., 2025. Stratifying prostate cancer: A comprehensive framework for staging and treatment planning classification. Procedia Computer Science 259, 356–365. URL: https://www.sciencedirect.com/science/article/pii/S1877050925010816, doi:https://doi.org/10.1016/j.procs.2025.03.337. sixth International Conference on Futuristic Trends in Networks and Computing Technologies (FTNCT06), held in Uttarakhand, India.

[19] Remya, S., Anjali, T., Sugumaran, V., 2024. A novel transfer learning framework for multimodal skin lesion analysis. IEEE Access 12, 50738–50754. doi:10.1109/ACCESS.2024.3385340.

[20] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S.R., Cole-Lewis, H., Neal, D., Rashid, Q.M., Schaekermann, M., Wang, A., Dash, D., Chen, J.H., Shah, N.H., Lachgar, S., Mansfield, P.A., Prakash, S., Green, B., Dominowska, E., Agüera y Arcas, B., Tomašev, N., Liu, Y., Wong, R., Semturs, C., Mahdavi, S.S., Barral, J.K., Webster, D.R., Corrado, G.S., Matias, Y., Azizi, S., Karthikesalingam, A., Natarajan, V., 2025. Toward expert-level medical question answering with large language models. Nature Medicine 31, 943–950. URL: https://doi.org/10.1038/s41591-024-03423-7, doi:10.1038/s41591-024-03423-7.

[21] Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., Paoli, J., Puig, S., Rosendahl, C., Soyer, P., Zalaudek, I., Kittler, H., 2020. Human–computer collaboration for skin cancer recognition. Nature Medicine 26, 1229–1234. doi:10.1038/s41591-020-0942-0.

[22] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I., 2017. Attention is all you need, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.