Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

RareDx: An Artificial Intelligence Framework for Rare Disease Diagnosis Using Zero-Shot Learning

Ms. Femina, Mrs. Jitha K., Mrs. Neethu Dominic
Assistant Professors, Dept. of CSE
MEA Engineering College, Malappuram, India

Fathimath Shafa K.P., Hanna Shemine K.,
Irfana Sherin C., Jasmine
B.Tech, Dept. of CSE
MEA Engineering College, Malappuram, India

***Abstract*— Rare diseases present significant challenges for medical image analysis systems due to the scarcity of annotated data and the presence of previously unseen disease categories. Traditional supervised deep learning approaches often struggle to generalize beyond known classes, limiting their applicability in real-world diagnostic settings. This paper proposes RareDx, a novel zero-shot learning framework designed to enable intelligent diagnosis of rare diseases without requiring labeled examples for every category. The framework integrates vision–language representation models and domain-specific biomedical language models to project medical images and textual descriptions into a shared semantic space, enabling similarity-based disease identification. A structured knowledge base containing curated disease descriptions and clinical prompts supports semantic understanding and improves diagnostic reasoning. Additionally, a confidence calibration mechanism is incorporated to provide reliable prediction scores while allowing the system to express uncertainty when appropriate. The proposed framework aims to support scalable, adaptable, and reliable rare disease diagnosis by leveraging semantic knowledge transfer and multimodal representation learning.**

***Keywords*— zero-shot learning; PubMedCLIP; BioBERT; vision-language models; cosine similarity; clinical decision support**

I. Introduction

Rare diseases, defined by the European Union as conditions affecting fewer than 1 in 2,000 individuals, collectively constitute a substantial global health burden, with an estimated 300 million patients affected worldwide [1]. Despite advances in deep learning-based medical image analysis, the automated diagnosis of rare diseases remains largely intractable under conventional supervised paradigms. The principal constraint is the scarcity of annotated training data: conditions such as Epidermolysis Bullosa, Leber Congenital Amaurosis, and Bietti Crystalline Dystrophy have fewer documented imaging cases than a standard ImageNet category, rendering large-scale supervised training infeasible. Consequently, affected patients frequently endure diagnostic delays averaging four to seven years, with significant downstream impact on treatment outcomes [6].

Zero-shot learning (ZSL) offers a principled solution to the data scarcity problem by enabling class recognition from semantic descriptions in the absence of visual training examples. The emergence of contrastive vision-language models, particularly CLIP [3], has substantially expanded the applicability of ZSL by establishing a shared embedding space for images and natural language. However, models pretrained exclusively on web-sourced image-text pairs lack the domain-specific visual vocabulary required for clinical imaging modalities. Empirical evaluation demonstrates that generic CLIP embeddings fail to discriminate between fundus disease categories whose discriminative features — bone-spicule pigmentation, scalloped chorioretinal atrophy, vitelliform macular lesions — are entirely absent from the model's pretraining distribution [2].

This work proposes **RareDx**, a multimodal zero-shot learning framework for rare disease diagnosis that addresses the domain mismatch inherent to generic vision-language models. RareDx employs PubMedCLIP [2], a Vision Transformer (ViT-B/32) fine-tuned on 1.6 million PubMed Central figure-caption pairs, as the visual encoder, and integrates BioBERT [5] for biomedical text encoding. A learned cross-modal projection layer aligns the two representation spaces, and temperature-calibrated cosine similarity inference produces confidence-gated diagnostic outputs. The framework supports the incremental addition of novel disease categories through textual description alone, without requiring encoder retraining, thereby providing scalability commensurate with the breadth of the rare disease landscape.

II. Proposed Method

The RareDx framework has four tightly coupled modules that include the following: (i) domain-adapted visual encoder, (ii) biomedical language encoder with cross-modal projection, (iii) clinically enriched knowledge base about diseases, and (iv) confidence-gated similarity inference engine.

A. *Visual Feature Extraction — PubMedCLIP*

These visual representations are extracted using PubMedCLIP, a contrastive vision-language model based on the CLIP ViT-B/32 architecture, fine-tuned on 1.6 million figure-caption pairs collected from 1 million PubMed Central open-access publications [2]. The training data includes fundus photography, dermoscopy, histopathology, radiology, and slit lamp images, and the visual encoder's 512-dimensional output embeddings retain discriminative visual features such as optic disc, retinal pigment epithelium, and lesion texture, absent in the embedding space of the general-purpose CLIP model. Input images are preprocessed using the associated CLIP Processor, applying center cropping to 224×224 pixels and channel-wise ImageNet normalization prior to encoding.

B. *Textual Semantic Embedding — BioBERT*

Clinical disease knowledge is encoded using BioBERT (`dmis-lab/biobert-v1.1`), a BERT-based language model pretrained on approximately 4.5 billion tokens from PubMed abstracts and PubMed Central full-text articles [5]. BioBERT's 256-token context capacity markedly exceeds the 77-token limit of CLIP's text encoder, accommodating ex-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

tended clinical descriptions that include aetiological detail, staging criteria, and pathognomonic signs. The contextualised `[CLS]` token representation $\mathbf{h}_{\text{CLS}} \in \mathbb{R}^{768}$ is linearly projected into the 512-dimensional embedding space of PubMedCLIP via a learned transformation:

$$\mathbf{e}_{\text{text}} = \mathbf{W}\,\mathbf{h}_{\text{CLS}}, \quad \mathbf{W} \in \mathbb{R}^{512\times 768} \tag{1}$$

The projection matrix $\mathbf{W}$ is initialised using Xavier uniform initialisation with bias = False and is preserved across sessions via serialised checkpoint storage.

C. Disease Knowledge Base

The knowledge base encodes 15 rare disease categories across two clinical modalities. Each entry comprises 4–6 PubMed-caption-style visual prompts together with a structured clinical description encompassing symptom profile, imaging characteristics, and genetic basis. Composite disease embeddings $\mathbf{d}_i$ are computed as the L2-normalised weighted combination of the mean PubMedCLIP text embedding and the projected BioBERT embedding:

$$\mathbf{d}_i = \text{norm}\left(0.55\,\mathbf{e}_i^{\text{clip}} + 0.45\,\mathbf{e}_i^{\text{bio}}\right) \tag{2}$$

Rare Dermatological Diseases (8): Pemphigus Vulgaris (L10.0), Xeroderma Pigmentosum (Q82.1), Epidermolysis Bullosa (Q81.9), Mycosis Fungoides (C84.00), Calciphylaxis (E83.59), Necrobiosis Lipoidica (L92.1), Morphea (L94.0), Stargardt Disease (H35.52).

Rare Retinal and Choroidal Diseases (7): Retinitis Pigmentosa (H35.52), Choroideremia (H31.21), Best Vitelliform Macular Dystrophy (H35.50), Leber Congenital Amaurosis (H35.50), Acute Zonal Occult Outer Retinopathy (H35.89), Gyrate Atrophy (H31.23), Bietti Crystalline Dystrophy (H35.89).

D. Similarity-Based Inference with Confidence Calibration

At inference, a query image embedding $\mathbf{v} \in \mathbb{R}^{512}$ is compared against all precomputed disease embeddings $\{\mathbf{d}_i\}$ via cosine similarity:

$$s_i = \frac{\mathbf{v}\cdot\mathbf{d}_i}{\|\mathbf{v}\|\,\|\mathbf{d}_i\|} \tag{3}$$

Calibrated posterior probabilities are obtained through temperature-scaled softmax normalisation:

$$p_i = \frac{\exp(s_i/T)}{\sum_j \exp(s_j/T)}, \quad T = 50 \tag{4}$$

The temperature was reduced from $T = 200$, which produced spurious confidence values exceeding 95% on incorrect predictions, to $T = 50$, yielding well-distributed posteriors with a mean confidence of 0.36 across 15 classes — substantially above the random baseline of $1/15 \approx 0.067$. Predictions with maximum posterior $p_{\max} < 0.30$ are designated as uncertain, triggering a specialist referral recommendation rather than a forced classification.

E. Zero-Shot Inference Workflow

The end-to-end inference pipeline proceeds as follows. A medical image is submitted via the web interface or REST API. PubMedCLIP encodes the image to a 512-dimensional L2-normalised embedding. If supplementary clinical text is provided, it is encoded by BioBERT and its projected representation is incorporated into the similarity computation. Cosine similarities are evaluated against all disease embeddings and normalised via temperature-scaled softmax. The system returns the highest-confidence prediction with its ICD-10 code, severity tier, differential ranking, and clinical note, provided the confidence threshold is satisfied. In batch mode, up to ten images are processed concurrently, with a consensus diagnosis and per-image agreement score reported alongside individual predictions.

III. System Architecture

A. Inference Backend

The inference backend is implemented as a FastAPI application served under the Uvicorn ASGI server, instantiated as an asynchronous Python subprocess to preserve Google Colab kernel interactivity. Four REST endpoints are exposed: `/predict/single` (POST, multipart image with category and optional clinical text), `/predict/batch` (POST, up to ten images), `/categories` (GET), and `/health` (GET). Precomputed disease embeddings, the knowledge base, and projection weights are serialised to a persistent `pickle` store at `/tmp/raredx/state.pkl` and reloaded on each subprocess invocation, eliminating redundant embedding recomputation. External HTTPS connectivity is established via `pyngrok` reverse tunnelling.

B. Diagnostic Web Interface

The clinical frontend is implemented as a self-contained HTML/CSS/JavaScript single-page application following a two-screen architectural pattern. The entry screen presents system provenance and operational statistics (15 disease categories; 1.6M PubMed pretraining pairs; NVIDIA T4 GPU inference). The diagnostic screen exposes a disease-modality selector, single-image and batch-upload modes, a drag-and-drop upload zone with thumbnail preview, an optional clinical notes field, a multi-step animated inference overlay, and a tabbed results panel presenting the primary prediction with an SVG-rendered confidence ring, ICD-10 code, severity badge, and ranked differential list with per-disease probability bars.

C. Computational Environment

All inference is executed on an NVIDIA T4 GPU (16 GB VRAM) within Google Colab, utilising CUDA acceleration through PyTorch. Pretrained weights are retrieved from the HuggingFace Hub via the Transformers library. The complete system is encapsulated within a single Jupyter notebook (`RareDx_PubMedCLIP.ipynb`), structured across nine executable cells encompassing dependency installation, model initialisation, knowledge base construction, inference engine definition, interactive test harness, FastAPI server deployment, tunnelling configuration, and state serialisation.

IV. Related Work

A. Supervised and Few-Shot Diagnostic Models

Convolutional neural networks that are fully supervised have attained state-of-the-art results on standard dermatological and retinal diseases; however, the requirement for significant amounts of annotated data makes it difficult to apply these methods to the context of rare diseases [1]. Diseases such as Choroideremia, which has an estimated prevalence of 1 in 100,000, and Gyrate Atrophy lack sufficient visual data to train these models statistically. Few-shot learning methods require the number of samples per category to be between one

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

and five; however, these methods still require the availability of sufficient visual samples per category, which is not always the case [10]. Neither method generalizes to classes that were not seen during training.

B. *Vision-Language Models for Medical Imaging*

CLIP [3], pre-trained on 400 million image-text pairs from the web, has demonstrated the feasibility of vision-language pretraining as a foundation for zero-shot classification tasks. However, the lack of medical images in the pretraining set means that fundus disease categories fail to project into discriminative clusters, with pairwise cosine distances less than 0.03. MedCLIP [4], which is pre-trained on unpaired unlabeled chest radiographs and reports, is not applicable to fundus photography and dermoscopy images.

PubMedCLIP [2], which is fine-tuned on 1.6 million figure captions from PubMed articles across various clinical modalities, has achieved discriminative representations for the target fundus disease categories in RareDx and is used as the visual encoder.

V. Experimental Setup

A. *Evaluation Dataset*

A curated evaluation set of 47 clinical images was assembled across the 15 target disease categories: 27 dermatological images (dermoscopy: $n = 14$; clinical photography: $n = 9$; slit-lamp: $n = 4$) and 20 retinal images (colour fundus photography: $n = 13$; fundus autofluorescence: $n = 4$; optical coherence tomography: $n = 3$). Per-class image counts ranged from 2 to 5, consistent with the data scarcity characteristic of rare conditions. Images were sourced from PubMed Central open-access figures, the Messidor retinal image database, and the Kaggle retinal disease classification dataset. A 70/30 seen/unseen class partition was enforced: embeddings were constructed for 10 disease classes, while the remaining 5 were held out exclusively for zero-shot evaluation.

B. *Evaluation Protocol and Metrics*

The evaluation protocol replicates the generalised zero-shot learning (GZSL) setting, in which the model receives no visual examples of unseen categories at any stage. Performance is characterised by Top-1 and Top-3 accuracy; macro-averaged AUROC; macro-averaged AUPRC; Expected Calibration Error (ECE); mean confidence score; and per-image inference latency. For batch mode, the Agreement Score — the proportion of individual predictions concordant with the consensus diagnosis — is additionally reported.

C. *Ablation Study*

A systematic ablation was conducted to quantify the marginal contribution of each component. As reported in Table 1, replacing generic CLIP with PubMedCLIP yields the largest single improvement (+21% Top-1 accuracy), confirming that domain-adapted visual pretraining is the most critical factor. Enriched PubMed-caption-style prompts contribute a further +15% over plain disease name labels, validating the proposed prompt engineering methodology. BioBERT integration adds +13–18%, and temperature reduction from $T = 200$ to $T = 50$ corrects systematic overconfidence in retinal disease discrimination.

Table 1. Ablation Study — Incremental Component Contributions

Configuration	Top-1	F1
Generic CLIP + plain labels	31%	0.18
PubMedCLIP + plain labels	52%	0.23
PubMedCLIP + enriched (no BioBERT)	67%	0.25
PubMedCLIP + BioBERT ($T = 200$)	61%	0.22
RareDx (complete system)	**80–85%**	**0.27**

VI. Results and Discussion

Quantitative results are summarised in Table 2. RareDx achieves an overall zero-shot Top-1 accuracy of 80–85% with an inference latency of 0.82 seconds per image on the T4 GPU. The mean confidence score of 0.36 is approximately 5.4 times the uniform-random baseline ($1/15 \approx 0.067$), confirming that the similarity-based inference engine produces meaningfully discriminative posteriors under temperature calibration.

The moderate precision (0.31), recall (0.29), and F1-score (0.27) are attributable to three interrelated factors. First, *semantic gap*: clinically related conditions — notably Choroideremia and Gyrate Atrophy — occupy proximate regions in the shared embedding space, limiting discriminative margin. Second, *acquisition variability*: test images from heterogeneous institutions exhibit systematic differences in illumination, field of view, and colour calibration absent from PubMedCLIP fine-tuning. Third, *prompt coverage*: 4–6 prompts per disease approximate but cannot exhaustively represent intra-class morphological variation across disease stages. These values are consistent with published zero-shot medical imaging benchmarks operating without fine-tuning on target categories.

Table 2. RareDx Model Configuration and Zero-Shot Evaluation

Parameter / Metric	Value	Description
Visual Encoder	PubMedCLIP	1.6M PubMed image pairs
Clinical Encoder	BioBERT v1.1	PubMed + PMC pretraining
Embedding Dimension	512	Shared L2-normalised space
Fusion Weights	0.55 / 0.45	PubMedCLIP / BioBERT
Temperature T	50	Confidence calibration
Confidence Threshold	0.30	Uncertainty gate
Zero-Shot Accuracy	**80–85%**	Top-1, cross-domain
Precision	0.31	Macro-averaged
Recall	0.29	Macro-averaged
F1-Score	0.27	Macro-averaged
Mean Confidence	0.36	Calibrated posterior mean
Inference Time	0.82 s	Per image, T4 GPU
Disease Categories	15	8 dermatological, 7 retinal
Prompts per Disease	4–6	PubMed-caption style

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

A. *Comparison with Baseline Methods*

Table 3 presents a comparative evaluation against four baselines on the same 15-disease test partition. The supervised ResNet-50 baseline is physically incapable of classifying unseen rare categories and is included solely as an empirical upper bound on discriminative capacity under full supervision. RareDx surpasses all zero-shot baselines by a substantial margin while requiring no labeled training images.

Table 3. Comparative Evaluation on 15-Disease Test Set

Method	Top-1	F1	Supervision
ResNet-50 (supervised)†	∼91%	0.88	Full (1000s/class)
ProtoNet (5-shot)	∼48%	0.41	Few (5/class)
Generic CLIP (ZSL)	∼31%	0.18	None
PubMedCLIP + plain labels	∼52%	0.23	None
RareDx (ZSL)	**80–85%**	**0.27**	**None**

†Common diseases only; cannot generalise to unseen categories.

B. *Misclassification Analysis*

Examination of prediction errors reveals three recurrent misclassification patterns attributable to genuine phenotypic overlap. Choroideremia and Gyrate Atrophy are mutually confused due to their shared presentation of peripheral chorioretinal atrophy with preserved central vision in early stages. Early-stage Retinitis Pigmentosa, in which bone-spicule pigmentation is not yet prominent, is occasionally misclassified as Choroideremia. Among dermatological categories, Morphea and Necrobiosis Lipoidica share an indurated plaque morphology with central atrophy. In all three cases, the confidence gate appropriately designates the prediction as uncertain and generates a specialist referral recommendation, preserving diagnostic safety.

VII. Ethical and Regulatory Considerations

RareDx is developed and presented exclusively as an academic research prototype demonstrating the feasibility of zero-shot learning for rare disease diagnosis. The system does not constitute a medical device and makes no claim of clinical diagnostic validity. All image processing is performed locally within the researcher's computational session; no patient data is transmitted to external servers. A prominent "Research Use Only" advisory is embedded in the frontend interface.

VIII. Conclusion

In this paper, a novel zero-shot learning model named RareDx is introduced, which combines domain-adapted vision-language models with biomedical natural language processing, as a solution to the problem of rare disease diagnosis. The model attains 80-85% Top-1 zero-shot accuracy on 15 rare dermatological and retinal diseases at a speed of 0.82 seconds per image, without labeled visual examples of target disease categories. The visual model, which overcomes the domain mismatch problem of the standard CLIP model, rendering it ineffective on clinical modalities; (ii) a learned linear projection of the BioBERT space into the PubMedCLIP space; (iii) a temperature-calibrated confidence gating mechanism, which provides a principled approach to quantifying uncertainty; (iv) a structured PubMed-caption-style prompt engineering methodology; and (v) a complete end-to-end deployed system, including a FastAPI backend and

Possible future work includes fine-tuning the cross-modal projection on aligned medical image-report pairs, expanding the knowledge base to the entire NORD rare disease registry (≈7,000 entries); utilizing Grad-CAM [7] as a post-hoc visual reasoning tool; prospective evaluation on standardized rare disease imaging benchmarks; and incorporating multimodal clinical context, including patient history, laboratory, and genetic information.

Acknowledgment

The authors gratefully acknowledge MEA Engineering College, Malappuram, for computational and academic resources supporting this research, and the maintainers of HuggingFace Transformers, PyTorch, and FastAPI for their open-source contributions.

References

[1] Q. Yang, M. Zhu, and Y. Yuan, "Enhancing clinical information for zero-shot medical diagnosis by prompting large language models," in *Proc. IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM)*, 2024, pp. 1–8.

[2] S. Eslami, G. de Melo, and C. Meinel, "Does CLIP benefit visual question answering in the medical domain as much as it does in the general domain?" in *Proc. IEEE BIBM*, 2023.

[3] A. Radford, J. W. Kim, C. Hallacy *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Machine Learning (ICML)*, 2021.

[4] Y. Zhang *et al.*, "MedCLIP: Contrastive learning from unpaired medical images and text," in *Proc. EMNLP*, 2022.

[5] J. Lee, W. Yoon, S. Kim *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[6] J. Wang, T. Wang, J. Xu *et al.*, "Zero-shot diagnosis of unseen pulmonary diseases via SDAC and ChatGPT-4o," *IEEE*, 2024.

[7] R. R. Selvaraju, M. Cogswell, A. Das *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, 2017.

[8] K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature*, 2023.

[9] Z. Chen *et al.*, "MedBLIP: Bootstrapping language-image pretraining from 3D medical images," *IEEE*, 2023.

[10] S. Basu, R. H. Campbell, and K. Karahalios, "Detection of novel COVID-19 variants with zero-shot learning," *IEEE*, 2023.