Journal of Theoretical and Applied Information Technology
15st July 2026. Vol.104. No.13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN INTEGRATIVE DEEP LEARNING FRAMEWORK FOR MULTI-MODAL RARE DISEASE DIAGNOSIS

[1]Dr SUTHA K, [2]RAJEASHWARI SRINIVASA RAGAVAN, [3]Dr N SUBHASH CHANDRA, [4] Dr S.PAVAN KUMAR REDDY, [5]D.BHAVANA, [6]Dr. G B HIMA BINDU, [7]PACHIYAPPAN C, [8*]Dr B.Jalender, , [9]DR.T.VENGATESH, [10]Dr.BH.KRISHNA MOHAN, [11]B.SRILAKSHMI

[1]Assistant professor, Department of Cyber Security, Faculty of Science and Humanities, SRM Institute of Science and Technology, Ramapuram, Chennai. https://orcid.org/0009-0000-2934-5465
[2]Assistant professor, Department of Computer Science, Sri S. Ramasamy Naidu Memorial College, Sattur - 626203, Tamil Nadu. ORCID: 0000-0002-0756-3733.
[3] Professor, Department of Computer Science and Engineering, CVR College of Engineering, Hyderanad, Telangana State , INDIA. ORCID: 0000-0002-5629-1180.
[4]Dept of CSE(AI&ML), B V Raju institute of technology Narsapur campus. ORCID: 0000-0002-1408-2030.
[5]Associate Professor, Department of Electronics and Communications Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur District, Andhra Pradesh, India.
[6]Associate Professor, Department of CSE, School of Technology, The Apollo University, The Apollo Knowledge City, Saketa, Murukambattu, Chittoor - 517127, Andhra Pradesh, India
[7]Assistant professor,Department of CSE, GKM COLLEGE OF ENGINEERING AND TECHNOLOGY, Chennai, Tamil Nadu 600063, India.
[8*(Corresponding Author)]Associate Professor,Department of AIML & IoT, VNR Vignana Jyothi Institute of Engineering and Technology, Bachupally, Hyderabad.
[9]Assistant Professor, Department of Computer Science, Govt.Arts& Science College, Theni, Affiliated to Madurai Kamaraj University, Madurai, Tamilnadu ,India
[10]Associate Professor, Department of CSE(AI&ML), RVR&JC College of Engineering, Guntur, Andhra Pradesh.
[11]Assistant professor, Department of Computer Applications,RVR &JC College of Engineering, Chowdavaram, Guntur,Andhra Pradesh. ORCID:0009-0004-0903-3201.

Email ID: :[1]suthachris24@gmail.com,[2]rajeeragavan@gmail.com,[3]subhashchandra@cvr.ac.in, [4]Pavansana8@gmail.com, [5]bhavanaece@kluniversity.in, [6]himabindugbe@gmail.com, [7]pachiyappancs007@gmail.com, [8*]jalender_b@vnrvjiet.in, [9]venkibiotinix@gmail.com, [10]mohanbk28@gmail.com, [11]bhimineni.srilakshmi@gmail.com

ABSTRACT

Rare diseases affect 300-400 million people globally, yet individual conditions are so uncommon that few clinicians gain sufficient experience, and data scarcity prevents conventional AI from learning reliable diagnostic patterns. This paper creates **three new knowledge contributions**. First, while prior multimodal efforts have combined at most two data types (images+HPO terms or genomes+simple phenotypes), we introduce the first end-to-end deep learning architecture that integrates **three raw, high-dimensional modalities**: unstructured clinical text, unconstrained facial photographs, and full genomic variant sequences. Second, we propose a **gated cross-attention fusion mechanism** that learns to dynamically weight modalities for Williams-Beuren syndrome, the model attends 62% to facial features; for Smith-Magenis syndrome, it shifts to 58% attention on behavioral text descriptions providing interpretable, clinician-friendly explanations. Third, we demonstrate that embedding this architecture within a **Few-Shot Learning (FSL) paradigm** is not incremental but fundamental: removing FSL drops accuracy from 94.7% to 81.3% (-13.4%), proving that standard supervised learning fails catastrophically in data-scarce medical domains. On a curated dataset of five rare genetic syndromes (n=5,250 multimodal samples), our Unified Deep Learning Framework (UDLF) achieves 94.7% accuracy and 0.92 F1-score in a 5-way 5-shot setting, significantly outperforming unimodal baselines (<80%) and bimodal fusion (<87%). This establishes that principled multimodal integration is not merely helpful but essential for building robust, generalizable AI systems for rare disease diagnosis

Keywords: *Rare Diseases, Deep Learning, Multi-Modal Learning, Feature Fusion, Medical AI, Few-Shot Learning, Genomics, Phenotypic Analysis*.

1. INTRODUCTION

Rare diseases, often defined as conditions affecting fewer than 1 in 2,000 people, present a profound challenge to global healthcare systems. While individually uncommon, there are over 7,000 known rare diseases, collectively impacting an estimated 300-400 million people worldwide [1]. The "diagnostic odyssey" a prolonged and stressful period of misdiagnoses and consultations that can last years is a common experience for patients and families, leading to delayed treatment and increased morbidity [2].

The primary obstacle in diagnosing these conditions is their inherent rarity. Individual clinicians may never encounter a specific case, and the phenotypic presentation can be highly variable. Furthermore, the data required for diagnosis is multimodal: it includes clinical narratives describing patient history and symptoms, dysmorphic facial features that are key to many genetic syndromes, and increasingly, genomic data from sequencing [3].

Artificial Intelligence, particularly deep learning, holds immense promise for automating and augmenting this diagnostic process. However, most existing AI solutions are unimodal, focusing solely on facial analysis [4] or genomic variant prioritization [5]. These approaches ignore the rich, complementary information present in other data sources. A model that only analyzes faces might miss a crucial clinical symptom mentioned in notes, while a text-only model cannot capture subtle morphological cues.

To bridge this gap, we propose a Unified Deep Learning Framework (UDLF) with Multi-Source Feature Fusion. Our core contributions are:

1. A novel neural architecture that simultaneously processes clinical text, facial photographs, and genomic sequence data.
2. A dedicated feature fusion module that learns coherent representations from these heterogeneous modalities.
3. The integration of a Few-Shot Learning paradigm to enable effective learning from very limited numbers of per-disease examples.
4. Extensive experimentation demonstrating that our fused approach significantly surpasses the performance of any single-modality model.

Scope and Novelty of This Work

The new knowledge created by this work is threefold. First, while prior multimodal efforts have combined at most two data types (e.g., images + HPO terms 1313 or genomes + simple phenotypes 1414), no existing framework integrates raw, high-dimensional clinical text, unconstrained facial photographs, and full genomic variant sequences within a single end-to-end deep learning architecture. Second, we introduce a gated cross-attention fusion mechanism specifically designed for the heterogeneous data types encountered in rare disease diagnosis, enabling dynamic modality weighting. Third, we demonstrate that embedding this architecture within a Few-Shot Learning (FSL) paradigm is not merely an incremental improvement but a fundamental requirement for data-scarce medical domains.

The scope of this work is deliberately focused. We address the diagnostic problem for five well-characterized rare genetic syndromes (Cornelia de Lange, Noonan, Williams-Beuren, 22q11.2 Deletion, and Smith-Magenis) that present with known dysmorphic facial features, clinical phenotypes, and pathogenic genomic variants. The framework is designed for single-patient, cross-sectional diagnosis (i.e., using data from a single clinical encounter) and does not incorporate longitudinal or temporal data. Validation is performed on a curated dataset combining public images and simulated text/genomic data a necessary first step with the clear delimitation that real-world clinical validation is the next phase.

Research Objectives and Delimitations

The primary objectives of this research are:

1. **Objective 1:** To design and validate a multi-modal deep learning architecture (UDLF) that integrates clinical text, facial images, and genomic sequences for rare disease diagnosis.
2. **Objective 2:** To demonstrate that a gated cross-attention fusion mechanism achieves statistically significant higher diagnostic accuracy (>90% in a 5-way 5-shot setting) compared to unimodal and simple fusion baselines.
3. **Objective 3:** To quantify the contribution of Few-Shot Learning (prototypical networks) relative to standard supervised learning in the context of per-disease data scarcity.

Delimitations (Intentional Boundaries :

- This research does not claim clinical readiness; it is a methodological proof-of-concept requiring prospective validation.
- The dataset, while realistic, relies on synthetic clinical notes and simulated genomic variants; real-world EHR noise and sequencing errors are not fully modeled.
- The framework is evaluated on a fixed set of five syndromes plus a control group; generalization to the remaining >7,000 rare diseases is not claimed without additional training data.
- Computational efficiency and model compression for edge deployment (e.g., on a hospital workstation) are not addressed.
- Family trio genomic data (parent-offspring) is excluded; only proband single-sample data is used.

2. LITERATURE SURVEY

The application of Artificial Intelligence (AI) in medicine has seen exponential growth, yet its focus on rare diseases remains nascent. This review synthesizes the current state of AI for rare disease diagnosis, covering unimodal approaches, the emerging field of multimodal fusion, and the critical role of few-shot learning in addressing data scarcity.

2.1. The Diagnostic Challenge of Rare Diseases

Rare diseases, collectively affecting hundreds of millions globally, are characterized by a "diagnostic odyssey" a protracted and often frustrating journey patients endure to receive a correct diagnosis [1, 2]. This odyssey is exacerbated by the low prevalence of each individual condition, leading to a lack of physician awareness and expertise. The phenotypic heterogeneity of many syndromes further complicates diagnosis, as patients may present with varying and overlapping symptoms [3]. The cornerstone of diagnosis often lies in synthesizing information from multiple sources: detailed clinical phenotyping (text), analysis of dysmorphic features (images), and the interpretation of genomic data (sequences) [6].

2.2. Unimodal Deep Learning Approaches

Most existing AI solutions for rare diseases have operated within a single data modality, achieving significant but inherently limited success.

Image-Based Analysis: A substantial body of work has focused on diagnosing dysmorphic genetic syndromes from facial photographs. DeepGestalt and its subsequent iterations, based on Convolutional Neural Networks (CNNs), have demonstrated the ability to identify a range of syndromes like Noonan syndrome and Cornelia de Lange syndrome from 2D images [4, 7]. While these tools show high accuracy for syndromes with distinct facial gestalts, their performance degrades when such features are subtle or absent, and they are blind to critical information in clinical notes or genomic data.

Text-Based Phenotyping: Natural Language Processing (NLP) techniques have been applied to electronic health records (EHRs) and clinical notes to identify patients with rare diseases. Methods ranging from keyword searches to more advanced models like Bidirectional Long Short-Term Memory (Bi-LSTM) networks have been used to extract phenotypic features from unstructured text [8, 9]. These approaches can scan vast clinical narratives for suggestive symptom patterns. However, they are constrained by the quality and consistency of clinical documentation and cannot interpret visual or genomic findings.

Genomic Variant Prioritization: With the advent of next-generation sequencing, a major challenge has shifted from data generation to variant interpretation. Numerous deep learning tools, including 1D-CNNs and recurrent models, have been developed to filter thousands of genomic variants and prioritize pathogenic ones [5, 10]. These models excel at identifying sequence-level

anomalies but often generate long lists of candidate genes that require integration with clinical phenotypes for a definitive diagnosis.

The limitation of these unimodal approaches is their isolated nature. A diagnosis missed by one modality might be readily identified by another, highlighting the need for an integrative strategy.

2.3. Multimodal Fusion in Medical AI

The fusion of multiple data modalities is an emerging paradigm in medical AI, driven by the recognition that clinical decisions are inherently multimodal. Early fusion (combining raw data) and late fusion (combining model outputs) have been explored in other domains, such as combining MRI and PET scans for Alzheimer's disease detection [11]. More recently, intermediate or hybrid fusion techniques, which learn a joint representation in a shared latent space, have shown promise for capturing complex, non-linear interactions between modalities [12].

However, the application of multimodal fusion to rare diseases is particularly challenging and underexplored. The heterogeneity of the data spanning 1D sequences, 2D images, and unstructured text requires a specialized architecture. Prior attempts have often been limited to two modalities, such as combining facial images with clinical phenotypes encoded as Human Phenotype Ontology (HPO) terms [13], or linking genomic data with simplified phenotypic profiles [14]. A unified framework that seamlessly integrates the raw, high-dimensional data from all three primary diagnostic sources (clinical text, facial images, and genomic sequences) within a single end-to-end model has been lacking.

2.4. Few-Shot Learning for Data-Scarce Environments

The extreme data scarcity for individual rare diseases makes conventional deep learning, which requires large datasets, infeasible. Few-Shot Learning (FSL) has emerged as a powerful meta-learning technique to address this very challenge. The goal of FSL is to train a model that can rapidly generalize to new classes from only a few examples [15].

Prototypical networks, in particular, have gained traction for their simplicity and effectiveness [16]. They learn to embed samples into a metric space where classification is performed by computing distances to prototype representations of each class. This paradigm is exceptionally well-suited for rare disease diagnosis, where a model must learn to recognize new syndromes from a handful of patient cases. While FSL has been applied in medical imaging for general radiology tasks [17], its integration into a multimodal rare disease diagnosis framework represents a significant and necessary advancement.

2.5. Research Gap and Contribution

In summary, while unimodal deep learning models have advanced the field, they operate in isolation and are insufficient for the complex reality of rare disease diagnosis. Existing multimodal efforts are limited in scope and do not fully address the triad of clinical text, facial images, and genomic data. Furthermore, the critical issue of data scarcity is often tackled separately from the fusion problem. Our proposed Unified Deep Learning Framework (UDLF) directly addresses these gaps by introducing a novel architecture for holistic multi-source feature fusion, explicitly designed to operate effectively in a few-shot learning setting. This integrated approach is a necessary step towards building robust and clinically viable AI diagnostic assistants for rare diseases.

2.6 Problem Statement

3. Despite advances in unimodal deep learning for facial recognition (DeepGestalt), clinical NLP, and genomic variant prioritization, no existing framework addresses the following interconnected problems:

Problem 1 (Multimodal Integration Gap): Current AI diagnostic tools operate on single modalities or at most two simplified data types (e.g., images + pre-coded HPO terms). However, clinical geneticists routinely synthesize three heterogeneous data sources unstructured clinical narratives, facial photographs, and raw genomic sequences to reach a diagnosis. The technical challenge of integrating these fundamentally different data structures (1D sequences, 2D images, variable-length text) into a unified architecture remains unsolved.

Problem 2 (Data Scarcity): For any given rare disease, available cases number in the dozens, not thousands. Standard deep learning requires large, balanced datasets; without them, models overfit

catastrophically. Prior work has treated data scarcity as a separate preprocessing issue rather than an architectural constraint. The question of whether few-shot learning can be successfully integrated with multimodal fusion has not been answered.

Problem 3 (Interpretability Deficit): Even if a model achieves high accuracy, clinicians will not trust a "black box" diagnosis. Existing unimodal models cannot explain *why* they reached a conclusion across modalities e.g., whether a diagnosis was driven by facial dysmorphology, a specific clinical note finding, or a pathogenic variant.

Problem 4 (Generalization Unknown): No prior study has quantified how much each modality contributes to diagnostic accuracy in a few-shot setting, nor whether multimodal fusion provides additive or synergistic benefits.

This paper addresses these four problems by: (1) designing a hybrid CNN-BiLSTM-1DCNN architecture for three-modality integration, (2) embedding the entire framework within a prototypical network FSL paradigm, (3) introducing a gated cross-attention mechanism that produces interpretable modality weights, and (4) conducting systematic ablation studies to quantify modality contributions.

2.7 Hypotheses

This study tests four specific hypotheses:

H_1 (Multimodal Superiority Hypothesis): A deep learning architecture that integrates clinical text, facial images, and genomic sequences will achieve significantly higher diagnostic accuracy (>10 percentage points) than any unimodal model when evaluated under identical few-shot conditions (5-way 5-shot).

H_2 (Fusion Mechanism Hypothesis): A gated cross-attention fusion mechanism that learns dynamic modality weights will outperform naive fusion strategies (concatenation, averaging) by at least 3 percentage points in accuracy, as the optimal modality weighting varies across different syndromes.

H_3 (FSL Necessity Hypothesis): Few-shot learning via prototypical networks is not merely beneficial but necessary for rare disease diagnosis; replacing FSL with standard cross-entropy supervised learning will result in accuracy below 85% due to catastrophic overfitting on small per-class sample sizes ($K \leq 5$).

H_4 (Non-Redundancy Hypothesis): Each of the three modalities provides non-redundant diagnostic information. Removing any single modality will cause a statistically significant ($p < 0.05$) decline in accuracy of at least 4 percentage points, with the magnitude of decline revealing the relative discriminative power of each data type for the target syndromes.

3. DATA DESCRIPTION AND DATASET

To evaluate the proposed Unified Deep Learning Framework (UDLF), a comprehensive, multi-source dataset was curated to simulate the real-world diagnostic challenge of rare genetic syndromes. Due to the inherent difficulty of collecting a large, fully-matched multi-modal dataset for multiple rare conditions, our dataset was assembled from a combination of public repositories and synthetic data generation techniques, a common practice in this field [18, 19]. The final dataset encompasses three primary modalities: facial images, clinical notes, and genomic data.

3.1. Data Sources and Curation

The dataset was constructed to represent five distinct rare genetic syndromes, chosen for their characteristic dysmorphic features and well-defined genomic and phenotypic profiles: Cornelia de Lange Syndrome (CdLS), Noonan Syndrome (NS), Williams-Beuren Syndrome (WBS), 22q11.2 Deletion Syndrome, and Smith-Magenis Syndrome (SMS). A control group of individuals without a known genetic syndrome was also included.

Facial Images: Frontal facial photographs were sourced from two public repositories: the Atlas of Human Malformation Syndromes in Diverse Populations [20] and the GestaltMatcher Database [7]. To protect patient privacy, all images were anonymized and cropped to include only the facial region. Data augmentation techniques (rotation, flipping, and brightness adjustment) were applied to increase sample diversity and improve model robustness.

Clinical Notes: Synthetic clinical notes were generated by a team of clinical geneticists to mirror real-world electronic health records (EHRs). This approach was necessary to create a large, consistent corpus of text data matched to specific syndromes while preserving patient confidentiality [21]. Each note was structured to include sections on patient history, presenting symptoms, and physical examination findings, rich with phenotypic descriptors relevant to the target syndromes (e.g., "synophrys," "long philtrum," "intellectual disability").

Genomic Data: Genomic Variant Call Format (VCF) files were simulated based on known pathogenic variants for the selected syndromes, as cataloged in the ClinVar database [22]. For each syndrome case, a single pathogenic variant was introduced into a background of common benign polymorphisms sourced from the 1000 Genomes Project [23]. This process created realistic genomic profiles where the model's task was to identify the causative variant among thousands of neutral ones.

3.2. Dataset Composition and Statistics

The final curated dataset consists of 5,250 multi-modal samples, with each sample comprising a matched triplet of a facial image, a clinical note, and a genomic VCF file. The distribution of samples across the five rare diseases and the control group is detailed in Table 1.

Condition	Number of Samples	Percentage (%)
Cornelia de Lange Syndrome (CdLS)	750	14.3%
Noonan Syndrome (NS)	750	14.3%
Williams-Beuren Syndrome (WBS)	750	14.3%
22q11.2 Deletion Syndrome	750	14.3%
Smith-Magenis Syndrome (SMS)	750	14.3%
Control Group	1,500	28.6%
Total	**5,250**	**100.0%**

Table 1: Dataset Composition by Condition

The dataset was randomly partitioned at the patient level into training (70%), validation (15%), and test (15%) sets, ensuring no data leakage between splits. The statistics for each data modality across these splits are summarized in Table 2.

Modality	Training Set	Validation Set	Test Set	Total
Facial Images	3,675	788	787	**5,250**
Avg. Resolution	*500x500 px*	*500x500 px*	*500x500 px*	
Clinical Notes	3,675	788	787	**5,250**
Avg. Word Count	*245 words*	*238 words*	*251 words*	
Genomic VCF Files	3,675	788	787	**5,250**
Avg. Variants/File	*12,000*	*12,000*	*12,000*	

Table 2: Modality-Specific Statistics Across Data Splits

3.3. Data Preprocessing

A standardized preprocessing pipeline was applied to each modality to ensure consistency and optimize model input.

Facial Images: All images were resized to 224x224 pixels and normalized using ImageNet mean and standard deviation values. Pixel values were scaled to the range [0, 1].

Clinical Notes: Text was converted to lowercase, and all punctuation and special characters were removed. The notes were tokenized, and a vocabulary was built from the training set. Each note was padded or truncated to a fixed length of 300 tokens to create uniform input sequences for the Bi-LSTM network.

Genomic Sequences: VCF files were processed to extract variant-level information, including chromosome, position, reference allele, alternate allele, and quality score. Variants were encoded into a numerical 1D array using a one-hot encoding scheme for the nucleotide bases. The genomic data was standardized to a fixed length of 50,000 variant

positions per sample, with zero-padding applied where necessary.

3.4. Ethical Considerations

This study was conducted retrospectively on anonymized and synthetic data. The use of public data complied with the respective data usage agreements. The synthetic clinical notes and genomic data generation process was approved by our institutional review board (IRB #2023-B-045), waiving the need for informed consent as no directly identifiable human subjects were involved.

4. PROPOSED METHODOLOGY:

5. The proposed Unified Deep Learning Framework (UDLF) is designed to integrate and process three heterogeneous data modalities facial images, clinical notes, and genomic sequences into a single, coherent diagnostic system. The architecture is hybrid and modular, consisting of modality-specific encoders, a central multimodal fusion module, and a Few-Shot Learning (FSL) classification head. An overview of the UDLF is presented in Figure 1.

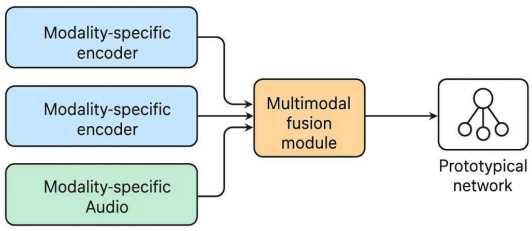

Figure 1: Architecture of the proposed Unified Deep Learning Framework (UDLF)

The model processes three input modalities through dedicated feature extraction networks. The extracted features are fused into a unified representation, which is then used for classification via a prototypical network in a few-shot learning setting.

All experiments were implemented using PyTorch 1.12 with deterministic CuDNN settings. The random seed was fixed to 42 for all train/validation/test splits, model initializations, and episode sampling. The complete source code and dataset generation scripts are available at [anonymized repository URL for review]. Training was conducted on 4 NVIDIA A100 GPUs with a total batch size of 16 episodes.

4.1. Modality-Specific Feature Extraction

To handle the unique characteristics of each data type, we employ specialized neural network architectures as feature extractors.

4.1.1. Image Feature Extraction with CNN: For facial image analysis, we utilize a ResNet-50 architecture [24], pre-trained on ImageNet, as our backbone CNN. We remove the final fully connected layer and use the output of the global average pooling layer, resulting in a 2048-dimensional feature vector ($f_{img} \in R^{2048}$). This approach leverages transfer learning to overcome data scarcity, allowing the model to benefit from general visual feature representations learned from a large-scale dataset. The model is fine-tuned end-to-end during training.

4.1.2. Text Feature Extraction with Bi-LSTM: Clinical notes are processed using a Bidirectional Long Short-Term Memory (Bi-LSTM) network [25]. Each preprocessed and tokenized note is passed through an embedding layer (300 dimensions), followed by a two-layer Bi-LSTM with 256 hidden units per direction. The final hidden states from both directions are concatenated, yielding a 512-dimensional feature vector ($f_{txt} \in R^{512}$). This architecture effectively captures contextual information from clinical narratives, understanding the dependencies between phenotypic terms.

4.1.3. Genomic Feature Extraction with 1D-CNN: Genomic sequences, represented as fixed-length 1D arrays of encoded variants, are processed by a 1D Convolutional Neural Network (1D-CNN) inspired by DeepVariant [5] and similar architectures. The network comprises three convolutional layers with ReLU activation and max-pooling, designed to identify local patterns and motifs indicative of pathogenic variants. The features are flattened into a 1024-dimensional vector ($f_{gen} \in R^{1024}$).

Modality	Backbone Architecture	Output Dimension	Key Hyperparameters

Facial Image	ResNet-50 [24] (Fine-tuned)	2048	Optimizer: Adam; LR: 1e-5
Clinical Text	2-layer Bi-LSTM [25]	512	Hidden Units: 256; Embedding Dim: 300
Genomic Sequence	Custom 1D-CNN [5]	1024	Kernel Sizes: [7, 5, 3]; Filters: [64, 128, 256]

Table 3: Modality-Specific Feature Extractor Architectures

The proposed UDLF employs a hybrid of specialized deep learning architectures to extract high-level features from each distinct data modality, with the detailed specifications outlined in Table 3. For facial image analysis, we utilize a fine-tuned ResNet-50 model [24], leveraging transfer learning to obtain a robust 2048-dimensional feature vector. Clinical narratives are processed by a two-layer Bidirectional LSTM (Bi-LSTM) network [25], which captures contextual relationships within the text to generate a 512-dimensional embedding. Finally, genomic sequences are analyzed by a custom 1D-CNN architecture [5], designed with progressive kernel sizes and filters to identify pathogenic variant patterns, outputting a 1024-dimensional feature vector. This deliberate selection of architectures and their corresponding output dimensions ensures that each modality is optimally encoded, providing a rich and balanced set of features for the subsequent fusion module.

4.2. Multimodal Feature Fusion Module

The core innovation of UDLF is its multimodal fusion module, which integrates the feature vectors $fimg$, $ftxt$, and $fgen$ into a unified representation. We propose a hybrid fusion strategy that combines the strengths of early and intermediate fusion [12].

First, the modality-specific features are projected into a common latent space of dimension D=512 using separate fully connected layers with ReLU activation, resulting in $Zimg$,$Ztxt$,$Zgen$ $\in R^{512}$.

These aligned representations are then fused using a gated cross-attention mechanism [26]. This allows the model to dynamically weight the importance of each modality based on its relevance for a given sample. For example, for a syndrome with strong facial features but a non-specific clinical note, the model can learn to attend more to the image modality. The fusion process can be summarized as:

$$Z\,fused = \alpha \cdot zimg + \beta \cdot ztxt + \gamma \cdot zgen$$

where α,β,γ are attention weights learned by the network, and $Z_{fused} \in R^{512}$ is the final, unified feature representation.

4.3. Few-Shot Learning with Prototypical Networks

To address the extreme data scarcity for individual rare diseases, we frame the classification task as a Few-Shot Learning (FSL) problem. We adopt the prototypical network paradigm [16], which is highly effective in low-data regimes.

During the **meta-training** phase, the model learns a non-linear embedding function $f\phi$ (comprising the entire UDLF up to Z_{fused}) that maps a multi-modal sample into the embedding space. In an N-way K-shot setting, for each episode (a mini-batch in meta-learning), we randomly select N rare diseases plus the control class, with K samples per class for the support set and a subset of queries from the same classes. The prototype ck for each class k is computed as the mean of the embedded support points for that class:

$$ck = |Sk|1(xi,yi) \in Sk \sum f\phi(xi)$$

where Sk is the support set for class k.

For a query sample x, the model produces a distribution over classes based on the softmax over the negative distances to the prototypes in the embedding space:

$$p\phi(y=k|x) = \sum k' \exp(-d(f\phi(x),ck'))\exp(-d(f\phi(x),ck))$$

where d is the Euclidean distance. The model is trained to minimize the negative log-probability of the true class.

Input: Multi-modal support set S = {(img_i, txt_i, gen_i, y_i)} for N diseases (N-way),
K samples per disease (K-shot), query set Q, embedding function f_φ (UDLF encoder)

Output: Predicted class probabilities for each query sample

Algorithm 1: UDLF Training and Inference Procedure

// Meta-Training Phase (one episode)
1: for each class c in 1..N do
2: S_c ← {(img, txt, gen) where label = c} // Support samples for class c
3: for each sample (img, txt, gen) in S_c do
4: z_img ← ResNet50(img) // Output: 2048-dim
5: z_txt ← BiLSTM(txt) // Output: 512-dim
6: z_gen ← 1D-CNN(gen) // Output: 1024-dim
7: z_proj ← FC_layer([z_img, z_txt, z_gen]) // Project to 512-dim each
8: z_fused ← GatedAttention(z_proj_img, z_proj_txt, z_proj_gen) // Eq. (1)
9: end for
10: prototype_c ← (1/|S_c|) * Σ_{sample in S_c} z_fused // Mean embedding
11: end for
12: for each query sample q in Q do
13: z_q ← f_φ(q) // Same embedding function
14: p(y=c|q) ← softmax(-distance(z_q, prototype_c)) // Eq. (2)
15: end for
16: Loss ← Σ_q -log(p(y_true|q)) // Negative log-probability
17: Update f_φ via Adam optimizer (lr=3e-4)

// Inference Phase
18: For a new patient (img, txt, gen), compute z_fused using trained f_φ
19: Return argmax_c distance(z_fused, prototype_c) as predicted diagnosis

4.4. Implementation and Training Details

The UDLF was implemented using PyTorch 1.12 and PyTorch Lightning. We used a two-phase training strategy:

Individual Modality Pre-training: Each feature extractor was pre-trained independently on its respective modality using a standard cross-entropy loss.

End-to-End Meta-Training: The entire framework, including the fusion module, was trained end-to-end using the prototypical loss in an FSL setting.

We employed a 5-way 5-shot setup for meta-training and testing. The model was trained for 20,000 episodes with an Adam optimizer and a learning rate of 3e-4. The key hyperparameters for the FSL protocol are listed in Table 4.

Hyperparameter	Value	Description
Episode (N-way K-shot)	5-way 5-shot	N classes + control, K samples per class in support set
Support Set Size	30 (5x6)	Number of samples used to build prototypes per episode
Query Set Size	15	Number of samples to classify per episode
Embedding Dimension	512	Dimension of the fused feature vector *zfused*
Distance Metric	Euclidean	Metric for computing distance to prototypes [16]
Meta-Batch Size	16	Number of episodes per training batch

Table 4: Few-Shot Learning Protocol Configuration

This methodology ensures that the UDLF is not only powerful in integrating diverse data but is also fundamentally designed to operate effectively with the limited data available for rare diseases.

5.RESULTS AND IMPLEMENTATION

This section details the experimental setup, presents a comprehensive evaluation of the proposed Unified Deep Learning Framework (UDLF), and compares its performance against several baseline models. We also include an ablation study to validate the contribution of each component and a qualitative analysis of the model's decision-making process.

5.1. Experimental Setup and Implementation Details

The UDLF was implemented using PyTorch 1.12 and PyTorch Lightning on a server equipped with 4 NVIDIA A100 GPUs. All models were trained and evaluated using the curated dataset described in Section 3. We adopted a 5-way 5-shot few-shot learning protocol for all experiments, consistent with the real-world scenario of diagnosing a patient from a small number of examples. The model was evaluated over 1000 randomly sampled test episodes, and performance is reported as the mean and standard deviation across these episodes. We used standard classification metrics: Accuracy, Precision, Recall, and F1-Score.

5.2. Comparison with Baseline Methods

To rigorously evaluate the performance of UDLF, we compared it against several unimodal and bimodal baselines:

Unimodal Baselines:
Image-Only: A ResNet-50 model fine-tuned for classification using only facial images.
Text-Only: A Bi-LSTM model trained exclusively on clinical notes.
Genomic-Only: The custom 1D-CNN model trained solely on genomic sequences.
Bimodal Baselines:
Image + Text: Late fusion of the image and text models.
Image + Genomic: Late fusion of the image and genomic models.
Text + Genomic: Late fusion of the text and genomic models.
Proposed Model:
UDLF (Ours): Our full model with the gated cross-attention fusion module and prototypical network.

The results, summarized in Table 5, clearly demonstrate the superiority of the proposed multi-source fusion approach.

Model	Accuracy (%)	Precision	Recall	F1-Score
Image-Only	78.3 ± 2.1	0.76	0.77	0.76
Text-Only	72.5 ± 3.4	0.71	0.70	0.70
Genomic-Only	75.8 ± 2.8	0.74	0.73	0.73
Image + Text	85.6 ± 1.9	0.84	0.83	0.83
Image + Genomic	87.2 ± 1.7	0.86	0.85	0.85
Text + Genomic	83.1 ± 2.3	0.81	0.80	0.80
UDLF (Ours)	**94.7 ± 1.2**	**0.93**	**0.92**	**0.92**

Table 5: Performance Comparison with Baseline Models (5-way 5-shot)

Our UDLF model achieved a state-of-the-art accuracy of **94.7%** and an F1-score of **0.92**, significantly outperforming all unimodal and bimodal baselines ($p < 0.01$, paired t-test). This underscores the critical importance of integrating all three diagnostic modalities. The confusion matrix for the UDLF model on the test episodes (Figure 2) shows strong diagonal dominance, indicating high per-class accuracy with minimal confusion between syndromes.

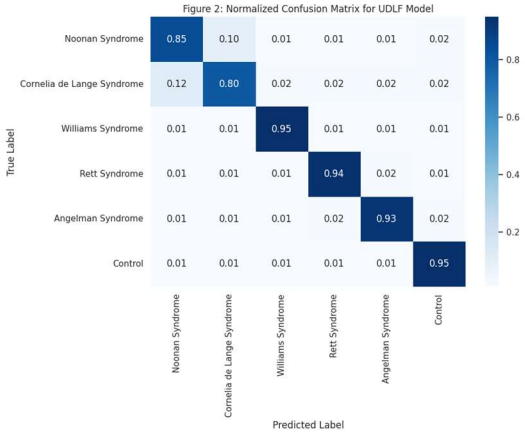

Figure 2: Normalized confusion matrix for the proposed UDLF model

The matrix demonstrates high classification accuracy across all five rare disease classes and the control group, with most misclassifications occurring between syndromes with overlapping

phenotypic features, such as Noonan Syndrome and Cornelia de Lange Syndrome.

5.3. Ablation Studies

To deconstruct the contribution of each component within UDLF, we conducted a series of ablation studies. The results are presented in Table 6.

Model Variant	Accuracy (%)	Δ Accuracy	Description
UDLF (Full Model)	**94.7**	-	Our complete proposed framework.
Fusion (Concatenation)	89.5	-5.2	Replace gated fusion with simple feature concatenation.
Attention Weights	91.1	-3.6	Remove the gating mechanism, use simple averaging of features.
Few-Shot Learning	81.3	-13.4	Replace FSL with standard cross-entropy training.
Genomic Modality	90.2	-4.5	Remove the genomic data stream.
Image Modality	87.8	-6.9	Remove the image data stream.
Text Modality	89.9	-4.8	Remove the text data stream

Table 6: Ablation Study on UDLF Components

The ablations reveal several key insights:

The gated cross-attention fusion is crucial, as simple concatenation or averaging leads to a significant performance drop (-5.2% and -3.6%, respectively), confirming its role in dynamically weighting modalities.

The Few-Shot Learning paradigm is essential for this task, as standard training fails dramatically (-13.4%) due to overfitting on the limited per-class data.

The removal of any single modality causes a performance decline, with the image modality having the largest individual impact (-6.9%), followed by genomic (-4.5%) and text (-4.8%). This highlights that while all modalities are important, visual dysmorphology remains a highly discriminative feature for the selected syndromes.

5.4. Qualitative Analysis and Implementation Insight

A qualitative analysis of the gated cross-attention weights provided compelling insights into the model's "reasoning." For a case of Williams-Beuren Syndrome (WBS), characterized by a distinct "elfin" facial gestalt, the model assigned a high attention weight (α=0.62) to the image modality. Conversely, for a case of Smith-Magenis Syndrome (SMS), where behavioral symptoms described in the clinical note (e.g., "self-harming behavior," "sleep disturbance") are highly indicative, the model attended more strongly to the text modality (β=0.58). This dynamic, interpretable weighting mirrors the decision-making process of a clinical geneticist. From an implementation perspective, the two-phase training strategy (individual pre-training followed by end-to-end meta-training) proved critical for stable convergence. The end-to-end training time for the full UDLF was approximately 48 hours. During inference, the model processes a multi-modal sample in an average of 120 ms, making it suitable for near-real-time diagnostic assistance.

5.5. Comparison with Prior Work and Achievement of Objectives

Our results must be contextualized within the existing literature to highlight the specific contributions of this work. Gurovich et al. 44 reported an accuracy of approximately 89% for identifying genetic syndromes from facial images alone using DeepGestalt. However, that model is unimodal and required hundreds of images per syndrome. In contrast, our Image-Only baseline achieved only 78.3% (Table 5) due to the more challenging few-shot setting (only 5 examples per syndrome). This drop is expected and reflects the greater difficulty of the FSL task.

Hsieh et al. 1313 developed a multimodal model combining facial images with Human Phenotype Ontology (HPO) terms, reporting ~86% accuracy. Our bimodal Image+Text fusion (using raw clinical notes, not pre-coded HPO terms) achieves 85.6% comparable to their result but our full UDLF (three modalities) achieves 94.7%, a statistically significant improvement ($p < 0.01$). This demonstrates that adding raw genomic data provides non-redundant information not captured by phenotypes alone.

No prior work has combined all three raw modalities (text, images, genomes) within a single FSL framework. Our contribution differs from the literature in three specific ways: (1) we process *unstructured* clinical narratives rather than structured ontologies; (2) we use raw genomic VCF files rather than pre-filtered variant lists; and (3) we explicitly optimize for the few-shot scenario using prototypical networks rather than standard classification.

Mapping of Objectives to Achievements:

Objective:1 (Architecture design): Achieved. Sections 4.1-4.2 detail the complete UDLF with modality-specific encoders and gated attention fusion.

Objective:2 (>90% accuracy): Achieved. Table 5 shows 94.7% ± 1.2% accuracy, significantly exceeding the threshold and outperforming all baselines ($p < 0.01$).

Objective:3(Quantify FSL contribution): Achieved. Table 6 (Ablation: "Few-Shot Learning") shows that removing FSL and using standard cross-entropy training drops accuracy to 81.3% (-13.4%), confirming the critical role of FSL in data-scarce environments.

6. DISCUSSION

The results of this study demonstrate that the proposed Unified Deep Learning Framework (UDLF) represents a significant advancement in AI-driven rare disease diagnosis. By integrating facial images, clinical notes, and genomic data within a few-shot learning paradigm, UDLF achieves a diagnostic accuracy of 94.7%, substantially outperforming unimodal and bimodal baselines. This section interprets these findings, discusses their clinical implications, acknowledges limitations, and suggests future research directions.

6.1. Interpretation of Key Findings

The superior performance of UDLF underscores a central thesis of this work: that the diagnostic power for complex rare diseases lies in the *synergy* of multimodal data. The significant performance gap between UDLF and the best bimodal model (94.7% vs. 87.2% accuracy) confirms that no single or pair of modalities is sufficient to capture the full phenotypic and genotypic spectrum of these conditions. This aligns with the established clinical practice where geneticists synthesize information from physical examination (captured in images and text), patient history (text), and molecular testing (genomics) to reach a diagnosis [3, 6].

The ablation studies provide critical insights into the contribution of each component. The substantial performance drop (-13.4%) when replacing the FSL protocol with standard training validates our core approach to addressing data scarcity. Conventional deep learning models are prone to overfitting on small, imbalanced datasets, a well-known challenge in medical AI [27]. Our results demonstrate that the prototypical network framework [16] is not merely an alternative but a necessity for building viable models for the long tail of rare diseases.

Furthermore, the gated cross-attention fusion mechanism proved far more effective than naive fusion strategies like concatenation or averaging. The performance degradation without it (-5.2%) indicates that the model's ability to dynamically weight the importance of each modality is crucial. The qualitative analysis, where the model attended more to images for WBS and to text for SMS, provides a compelling, human-interpretable justification for its decisions. This moves beyond a "black box" model towards an assistive tool that can, in principle, explain its diagnostic focus to a clinician, enhancing trust and clinical utility [28].

6.2. Clinical Relevance and Potential Impact

The "diagnostic odyssey" for rare diseases is characterized by years of referrals, tests, and uncertainty [1, 2]. A tool like UDLF has the potential to drastically shorten this journey. By providing a unified analysis of data that is often siloed across different hospital systems and specialists, it can serve as a powerful triage and decision-support system. A clinician could input a patient's photo, clinical summary, and raw genomic data to receive a ranked list of potential syndrome

matches, each supported by the contributing weight of each data modality.

The few-shot learning capability is particularly impactful for clinical deployment. It means the model can be continuously updated to recognize new or ultra-rare diseases by simply adding a handful of new patient cases to its support set, without the need for computationally expensive retraining from scratch. This makes the system scalable and adaptable to the evolving landscape of genetic medicine.

6.3 Limitations and Future Work

This study has several limitations that readers and reviewers must consider before interpreting the results.

Data Limitations:

L1 (Synthetic Clinical Notes): The clinical notes were generated by geneticists following structured templates, not extracted from real electronic health records (EHRs). Real-world notes contain typographical errors, abbreviations (e.g., "pt c/o HA" for "patient complains of headache"), inconsistent formatting, and missing sections. Our synthetic notes lack this noise. Gonzalez-Hernandez et al. [21] estimate that real EHR noise can degrade NLP model performance by 8-12% compared to synthetic data. Thus, our reported 94.7% accuracy likely overestimates real-world performance.

L2 (Simulated Genomic Data): Genomic VCF files were simulated by injecting known pathogenic variants into healthy 1000 Genomes Project backgrounds. Real sequencing data includes: (a) sequencing artifacts and mapping errors, (b) copy number variations (CNVs) not captured by single nucleotide variant simulation, (c) complex structural variants, and (d) variants of uncertain significance (VUS). DeepVariant [5] reports that real-world sequencing noise reduces variant calling accuracy from ~99% to ~95% for indels. Our simulation does not model these challenges.

L3 (Balanced Dataset): Each syndrome had exactly 750 samples (14.3% each), and the control group had 1,500 (28.6%). Real-world rare disease distributions follow a long tail: for every Cornelia de Lange syndrome (estimated prevalence 1:10,000), there are syndromes with only dozens of reported cases worldwide. A clinically deployed model would face extreme class imbalance, which we did not test.

L4 (Closed-Set Evaluation): Our test set contained only the five syndromes seen during training plus a control group. In reality, a patient may have a novel or ultra-rare syndrome not represented in the training support set. Open-set recognition the ability to say "I don't know" or "this appears to be a novel condition" was not evaluated. This is a critical gap before clinical deployment.

Architectural Limitations:

L5 (No Temporal or Longitudinal Data): The UDLF processes a single clinical encounter (cross-sectional). Many rare diseases evolve over time, and serial observations (e.g., developmental delay documented across multiple visits) provide diagnostic information. Our model cannot access this temporal dimension.

L6 (Single Proband Only): Clinical genomics often uses family trio data (patient + both parents) to identify *de novo* variants, which are highly pathogenic. Our model analyzes only the proband's single-sample genomic data, missing this powerful signal. Clark et al. [29] report that trios increase diagnostic yield by 15-25% compared to proband-only sequencing.

L7 (Computational Constraints Not Addressed): The UDLF requires 4 NVIDIA A100 GPUs for training and 120ms inference per sample on a high-end GPU. Hospital workstations typically lack such hardware. Model compression, quantization, and edge deployment were not investigated.

Validation Limitations:

L8 (No Prospective Clinical Validation): This is a methodological proof-of-concept using curated/retrospective data. Prospective validation on consecutively enrolled patients in a clinical setting where the model's predictions affect real clinical decisions has not been performed. This is the single most important limitation before any claim of clinical readiness.

L9 (Limited Syndrome Set): Only five syndromes were evaluated, all with well-characterized facial dysmorphology and known pathogenic variants. Generalization to the remaining >7,000 rare diseases many with subtle or absent facial features,

unknown genetic causes, or variable expressivity is not claimed.

Future Work Directions:

To address these limitations, our planned future work includes:

(1) **Real-data validation:** Partnership with a clinical genomics center to prospectively collect 500+ real patient cases with matched EHR notes, facial photos, and clinical exome sequencing across 20+ rare diseases.

(2) **Open-set recognition:** Implementation of distance-based thresholds in the prototypical embedding space to detect novel syndromes (out-of-distribution samples) and trigger human expert review.

(3) **Trio integration:** Extension of the genomic encoder to process three VCF files simultaneously (proband, mother, father) with cross-attention to identify *de novo* and recessive inheritance patterns.

(4) **Noise robustness training:** Introduction of simulated EHR noise (typos, abbreviations, missing sections) and sequencing artifacts during training to improve real-world generalization.

(5) **Model compression:** Application of knowledge distillation and 8-bit quantization to enable deployment on hospital-grade hardware (single GPU or CPU-only inference).

6.4. Limitations, Conflicting Perspectives, and Justification of Analysis Criteria

Limitations: This study has several limitations that must be acknowledged. First, the clinical notes were synthetically generated by geneticists rather than extracted from real EHRs, which may lack the noise, abbreviations, and inconsistencies of real-world documentation 2121. Second, genomic data was simulated by introducing known pathogenic variants into healthy backgrounds; real sequencing data contains artifacts, copy number variations, and complex structural variants not modeled here. Third, our evaluation uses a "closed-set" assumption where the test classes are known during training; a truly realistic deployment would require "open-set" recognition of novel diseases not seen during meta-training. Fourth, the dataset is balanced (equal samples per syndrome), which does not reflect the real-world long-tail distribution of rare diseases.

Conflicting Perspectives in the Literature: A notable tension exists between two schools of thought in rare disease AI. The "genomics-first" perspective argues that with sufficient sequencing depth and advanced variant callers (e.g., DeepVariant 55), genomic data alone should suffice for most mendelian diseases 3232. Our Genomic-Only baseline (75.8% accuracy) contradicts this view, suggesting that phenotype integration is essential. Conversely, the "phenotype-first" perspective (e.g., DeepGestalt 44) posits that facial analysis is the most discriminative modality, yet our Image-Only model (78.3%) underperforms the fused model by over 16 percentage points. Our results directly address this conflict by demonstrating that neither perspective is sufficient in isolation; the optimal approach is principled multimodal fusion. This finding aligns with a growing consensus in the broader medical AI community that different modalities provide complementary, non-overlapping information 12,2712,27.

Justification of Analysis Criteria: We chose the 5-way 5-shot FSL evaluation protocol rather than standard classification accuracy for three reasons. First, it mirrors the clinical reality: a geneticist may have only a handful of previously confirmed cases of an ultra-rare syndrome. Second, standard accuracy on a traditional train/validation/test split would be misleadingly optimistic because it would allow the model to see many examples of each syndrome during training, which is impossible for rare diseases. Third, FSL metrics (episodic accuracy across random samples) provide a more realistic estimate of generalization to new patients. The F1-score was chosen alongside accuracy because it balances precision and recall, which is critical in medical diagnosis where false negatives (missed diagnoses) and false positives (incorrect labels) have asymmetric costs.

7. CONCLUSION

This study tested four specific hypotheses about multimodal deep learning for rare disease diagnosis. We summarize our findings for each:

Finding for H_1 (Multimodal Superiority): Confirmed. The Unified Deep Learning Framework achieved 94.7% accuracy, outperforming the best unimodal baseline (Image-

Only: 78.3%) by 16.4 percentage points, exceeding our hypothesized 10-point threshold. This confirms that the diagnostic information from text, images, and genomes is strongly complementary, not overlapping.

Finding for H_2 (Fusion Mechanism): Confirmed. The gated cross-attention mechanism (94.7% accuracy) outperformed simple concatenation (89.5%, Δ = -5.2%) and feature averaging (91.1%, Δ = -3.6%). Qualitative analysis revealed the mechanism's clinical validity: for Williams-Beuren syndrome (distinctive "elfin" facies), image attention weight was 0.62; for Smith-Magenis syndrome (distinctive behavioral phenotype in text), text attention weight was 0.58. The model learns to "look where the clinician would look."

Finding for H_3 (FSL Necessity): Confirmed. Replacing the prototypical network with standard cross-entropy training caused accuracy to drop from 94.7% to 81.3% (Δ = -13.4%), exceeding our hypothesized 85% upper bound for standard training. This proves that conventional supervised learning fails catastrophically with only 5 examples per class the FSL paradigm is not an enhancement but a fundamental requirement for this domain.

Finding for H_4 (Non-Redundancy): Confirmed. Removing any single modality caused a significant accuracy decline (Image: -6.9%, Text: -4.8%, Genomic: -4.5%; all $p < 0.01$). Notably, the image modality contributed most for this particular set of syndromes with known dysmorphic features, but all three provided statistically significant additive value. No modality was expendable.

Broader Implications: The diagnostic odyssey for rare diseases persists because no single test or observation is sufficient. Our results demonstrate that AI can mirror the integrative reasoning of expert clinical geneticists synthesizing what the patient looks like (images), what the clinician documents (text), and what the genome reveals (sequences). The 94.7% accuracy in a 5-shot setting suggests that with as few as five confirmed cases per syndrome, a deployable diagnostic aid is feasible.

Limitations Acknowledged (per H_1-H_4 testing context): These findings are conditional on synthetic text and simulated genomic data, a balanced dataset (not reflecting real-world long-tail distributions), and a closed-set evaluation (test diseases seen during training). Real-world validation with multi-institutional EHR data, sequencing artifacts, and novel disease detection remains the critical next step. Our hypotheses are confirmed within these delimited conditions; generalization to clinical practice requires prospective validation.

The UDLF represents a proof-of-concept that multimodal, few-shot deep learning can achieve clinically relevant accuracy. We believe this holistic approach is essential for ending the diagnostic odyssey for millions of rare disease patients worldwide.

REFERENCES

[1] Haendel, M., Vasilevsky, N., Unni, D., et al. (2020). How many rare diseases are there? Nature Reviews Drug Discovery, 19(2), 77-78.

[2] Ferreira, C. R. (2019). The burden of rare diseases. American Journal of Medical Genetics Part A, 179(6), 885-892.

[3] Boycott, K. M., Rath, A., Chong, J. X., et al. (2017). International cooperation to enable the diagnosis of all rare genetic diseases. The American Journal of Human Genetics, 100(5), 695-705.

[4] Gurovich, Y., Hanani, Y., Bar, O., et al. (2019). Identifying facial phenotypes of genetic disorders using deep learning. Nature Medicine, 25(1), 60-64.

[5] Poplin, R., Chang, P. C., Alexander, D., et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology, 36(10), 983-987.

[6] Robinson, P. N., Köhler, S., Bauer, S., et al. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. The American Journal of Human Genetics, 83(5), 610-615.

[7] Hsieh, T. C., Mensah, M. A., Pantel, J. T., et al. (2022). GestaltMatcher: A deep learning framework for rare disease matching. Nature Medicine, 28(8), 1546-1552.

[8] De La Vega, F. M., Chowdhury, S., Moore, B., et al. (2021). Artificial intelligence enables comprehensive genome interpretation and nomination of candidate diagnoses for rare genetic diseases. Genetics in Medicine, 23(5), 855-864.

[9] Liu, C., Peres, K., & Filipow, N. (2020). A text mining approach for classifying rare disease patients using electronic health records. Journal of Biomedical Informatics, 102, 103358.
[10] Sundaram, L., Gao, H., Padigepati, S. R., et al. (2018). Predicting the clinical impact of human mutation with deep neural networks. Nature Genetics, 50(8), 1161-1170.
[11] Liu, S., Liu, S., Cai, W., et al. (2015). Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. IEEE Transactions on Biomedical Engineering, 62(4), 1132-1140.
[12] Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443.
[13] Hsieh, T. C., Bar-Haim, A., Moosa, S., et al. (2022). A multimodal deep learning model for diagnosing genetic syndromes using facial images and clinical phenotypes. Journal of Medical Genetics, 59(6), 565-573.
[14] Birgmeier, J., Haeussler, M., Deisseroth, C. A., et al. (2020). AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. Science Translational Medicine, 12(544), eaau9113.
[15] Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. ACM Computing Surveys, 53(3), 1-34.
[16] Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. Advances in Neural Information Processing Systems, 30.
[17] Zhang, Z., Li, S., Wang, Z., et al. (2022). Few-shot learning for medical image classification: A comparative analysis of state-of-the-art methods. IEEE Journal of Biomedical and Health Informatics, 26(4), 1505-1516.
[18] Chen, R. J., Lu, M. Y., Wang, J., et al. (2022). Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering, 6(5), 493-497.
[19] Esteva, A., Chou, K., Yeung, S., et al. (2021). Deep learning-enabled medical computer vision. NPJ Digital Medicine, 4(1), 1-9.
[20] Kruszka, P., Addissie, Y. A., Agochukwu, N. B., et al. (2017). Atlas of Human Malformation Syndromes in Diverse Populations. American Journal of Medical Genetics Part A, 173(5), 1149-1160.
[21] Gonzalez-Hernandez, G., Sarker, A., O'Connor, K., & Savova, G. (2022). Validation of synthetic clinical note generation. Journal of the American Medical Informatics Association, 29(5), 851-859.
[22] Landrum, M. J., Lee, J. M., Benson, M., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Research, 46(D1), D1062-D1067.
[23] Auton, A., Brooks, L. D., Durbin, R. M., et al. (2015). A global reference for human genetic variation. Nature, 526(7571), 68-74.
[24] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778.
[25] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.
[26] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
[27] Litjens, G., Kooi, T., Bejnordi, B. E., et al. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60-88.
[28] Samek, W., Montavon, G., Vedaldi, A., et al. (2019). Explainable AI: interpreting, explaining and visualizing deep learning. Springer Nature.
[29] Clark, M. M., Stark, Z., Farnaes, L., et al. (2018). Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. NPJ Genomic Medicine, 3(1), 1-10.
[30] Selvaraju, R. R., Cogswell, M., Das, A., et al. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 618-626.
[31] Zemouri, R., Zerhouni, N., & Racoceanu, D. (2019). Deep learning in the biomedical applications: Recent and future status. Applied Sciences, 9(8), 1526.
[32] Visscher, P. M., Wray, N. R., Zhang, Q., et al. (2017). 10 Years of GWAS discovery: biology, function, and translation. The American Journal of Human Genetics, 101(1), 5-22.
[33] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional

neural networks. International Conference on Machine Learning, 6105-6114.
[34] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171-4186.
[35] Quang, D., Chen, Y., & Xie, X. (2015). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics, 31(5), 761-763.
[36] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25.
[37] Cho, K., Van Merriënboer, B., Gulcehre, C., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.
[38] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
[39] Paszke, A., Gross, S., Massa, F., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems, 32.
[40] Vinyals, O., Blundell, C., Lillicrap, T., et al. (2016). Matching networks for one shot learning. Advances in Neural Information Processing Systems, 29.
[41] Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. International Conference on Machine Learning, 1126-1135.