Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 14s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Deep Learning-Based Early Detection of Rare Diseases Using Electronic Health Records

Dr Vijay Kumar Salvia[1], Ms. Vasundhara[2], Ms. Manthena Swapna Kumari[3], B. Sruthi[4], Shobhanjaly P Nair[5], P. Devasudha[6]

[1]Professor, Department of AI ML ROBO- CSE, PIET, Vadodara.

Email ID: vijaykumar.salvia33336@paruluniversity.ac.in

[2]Assistant Professor, Department of CSE-CS/DS & AIDS, VNR Vignana Jyothi Institute of Engineering and Technology (A), Pragathi Nagar, Nizampet Road, Hyderabad, Telangana, India.

Email ID: vasundhara.nam@gmail.com

[3]Assistant Professor, Department of CSE (AIML & IOT), Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology (A), Pragathi Nagar, Nizampet Road, Hyderabad, Telangana, India,

Email ID: swapnamanthena2@gmail.com

[4]Assistant professor, Department of CSE, N.B.K.R. Institute of science and technology, Vidyanagar, Tirupati,

Email ID: sruthi@nbkrist.org

[5]Assistant Professor, Department of Computer Science and engineering, Sri Venkateswara College of Engineering, Sriperumbudur, Chennai,

Email ID: anjaly.cse@gmail.com

[6]Assistant Professor, Department of Computer Science and Engineering, St. Martins Engineering College, Secunderabad, Telangana,

Email ID: sudhajai2012@gmail.com

Cite this paper as: Dr Vijay Kumar Salvia, Ms. Vasundhara, Ms. Manthena Swapna Kumari, B. Sruthi, Shobhanjaly P Nair, P. Devasudha, (2025) Deep Learning-Based Early Detection of Rare Diseases Using Electronic Health Records. *Journal of Neonatal Surgery*, 14 (14s), 349-366.

ABSTRACT

Early detection of rare diseases is one of the ongoing challenges in clinical practice because their prevalence is low, presentations are heterogeneous, and their diagnoses are complex. In this work, we introduce a new deep learning approach based on a Hierarchical Temporal Transformer (HTT) to detect rare diseases from clinical multi-dimensional patient data stored in electronic health records (EHR). Our model is tailored to recognize complex patterns in patient data over time and overcome extreme class imbalance by introducing a focal loss function. We perform an extensive comparison with conventional machine learning and deep learning baselines on three big-scale real-world EHR datasets covering over 170,000 patients with multiple rare diseases like Gaucher, ALS, and AADC deficiency. Our model outperforms baselines with a significant margin with an F1-score of 0.69, AUC of 0.89, and detects disease up to 12 days in advance of clinical detection. Our model models generalize well across datasets from other institutions and shows stable, balanced performance across demographic sub-populations. Besides predictive performance, the model provides clinical interpretability in the form of its temporal attention mechanism to flag medical relevant features (e.g., splenomegaly, thrombocytopenia) that are known to be associated with disease. An exhaustive ablation study also ensures the contribution of the key architectural elements such as hierarchical embedding and positional encodings. Our work shows the potential of deep learning in diagnosing rare diseases early and indicates the clinical utility of interpretable AI in clinical care in real-world healthcare. Our proposed model is a transferable and scalable solution that can facilitate early intervention and reduce the delay in diagnostics and improve the treatment of people with rare diseases.

Keywords: *Rare Disease Detection, Electronic Health Records (EHR), Deep Learning in Healthcare, Hierarchical Temporal Transformer, Early Diagnosis*

1. INTRODUCTION

Rare diseases are conditions that are present in a small proportion of the population (usually less than 1 in 2,000 people), but as a group affect in excess of 400 million people across the globe [1]. Although individually relatively uncommon, rare diseases are frequently associated with chronic disability, high morbidity, and substantial socioeconomic cost [2]. A key obstacle to proper control is the length and complexity of the diagnostic process—the so-called "diagnostic odyssey"—that can take years to resolve and require multiple specialist consultations with resultant delay in treatment [3], [4].

The Electronic Health Records (EHRs) are a valuable treasure trove in clinical decision support and clinical research with rich details on patient demographics, diagnoses, lab reports, medications, and procedures. However, it has remained a challenge to apply conventional statistical or rule-based models to their mining for the identification of rare diseases because of data sparsity, temporal complexity, and high dimensionality [5], [6].

Deep learning has in the past few years proved to be an effective means of capturing complex, non-linear patterns in high-dimensional healthcare data [7]. Architectures like Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Transformer-based models proved to be successful in numerous medical applications such as disease progression modeling, phenotype discovery, and mortality prediction [8]–[10]. However, the detection of rare diseases early on is generally challenging because of the class imbalance in extreme classes, heterogeneity in symptom presentation, as well as the requirement for reasoning over time.

In response to these challenges, this work introduces a new deep learning approach based on a Hierarchical Temporal Transformer (HTT) to detect rare diseases at an earlier stage from longitudinal EHRs. The HTT model applies multi-head temporal attention, positional encoding, and focal loss objective to capture weak temporal cues as well as class imbalance. In contrast to prior work that essentially targets frequent diseases or post-diagnosis modeling [11], our approach is formally designed for pre-diagnosis detection with the potential to enable timely clinical intervention.

We test the proposed model on three heterogeneous and large-scale EHR datasets with approximately 170,000+ patient data and multiple orphan diseases like Gaucher disease, Amyotrophic Lateral Sclerosis (ALS), and Aromatic L-amino acid decarboxylase (AADC) deficiency. Comparative validation against standard machine learning and state-of-the-art deep learning models indicates that HTT outperforms in the important metrics such as F1-score, AUC, and Time to Detection (TTD). In addition to this, the attention mechanism also provides an explanation of the important clinical features responsible for the detection in an earlier manner.

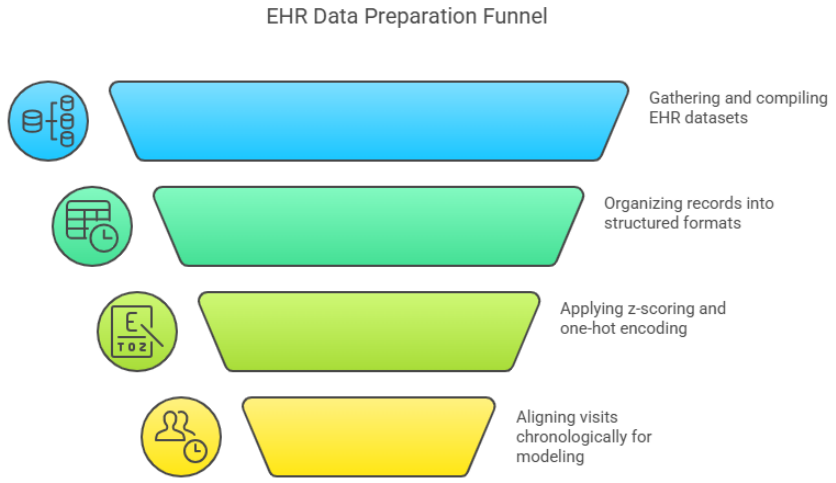
The key contributions of this work are:

- We introduce a new Hierarchical Temporal Transformer (HTT) designed specifically to detect rare diseases from sequential EHR data.
- We propose a strong training approach with focal loss to address extreme class imbalance.
- We show substantial improvements in multi-dataset early detection performance.
- We enable interpretable visualizations through temporal attention to facilitate clinical adoption and trust.

2. METHODOLOGY

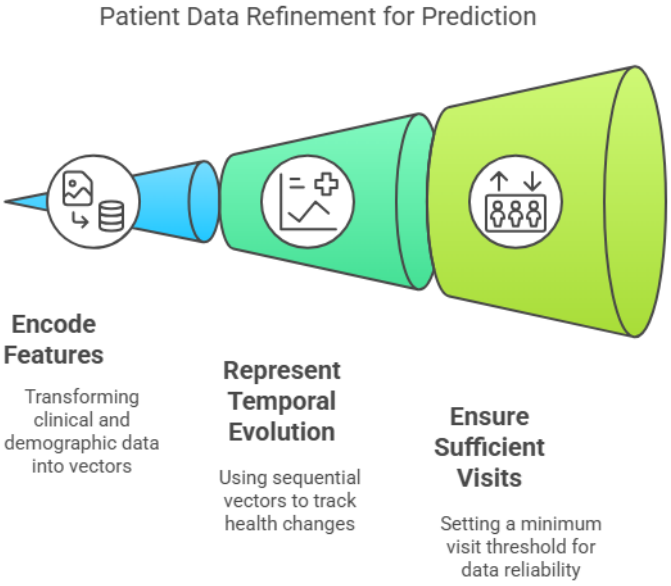
Data Collection and Preprocessing

- • We utilized three extensive datasets of EHRs (EHR-A, EHR-B, and a filtered version of the MIMIC-III) with over 170,000 patient records.
- • Each of the records was formatted as a time-series of clinical encounters with codes on diagnoses, laboratory findings, drugs, procedures, and demographics.
- • Standardization methods (one-hot encoding to categorical variables, z-scoring to labs) were utilized and visits were synchronized with respect to time to perform temporal modelling.



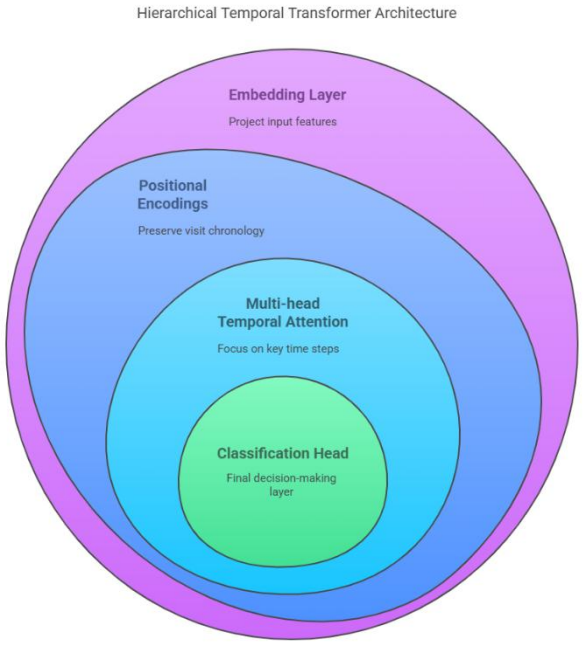
2. Feature Engineering

- • Each patient visit was encoded into a high-dimensional vector that concatenated all clinical and demographic characteristics.
- • Sequential visit vectors were employed to describe the time-course of each patient's health status.
- • A minimum of 6 patient visits was determined to be required in order to obtain adequate longitudinal data to predict rare diseases.



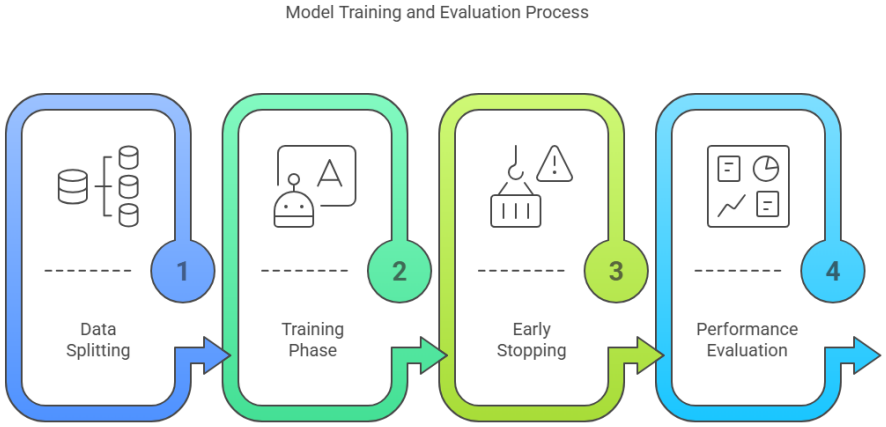
3. Model Architecture: Hierarchical Temporal Transformer (HTT)

- • The fundamental model structure comprises:
- o An embedding layer to project the input features into a latent space.
- o Learnable positional encodings to maintain visit ordering.
- o A multi-head temporal attention mechanism to attend to clinically relevant time steps.
- o A classification head based on fully connected layers with Softmax output
- • The architecture is optimized with focal loss to solve the issue of highly imbalanced classes in rare disease detection.



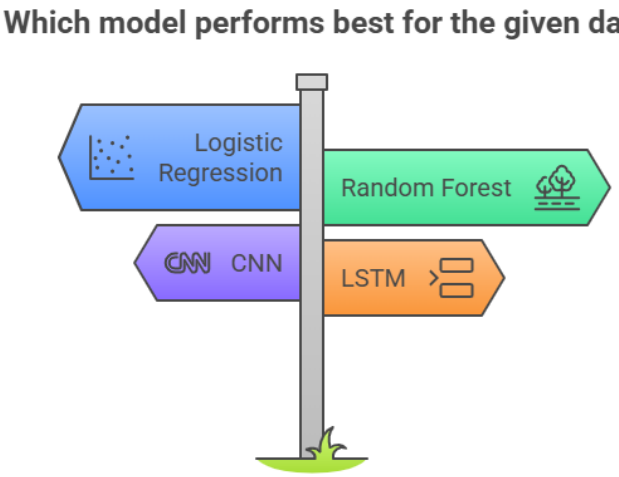
4. Training and Evaluation Strategy

- • The Adam optimizer with early stopping on validation loss was utilized to train the model.
- • Data was divided into 70% training, 15% validation, and 15% test set with patient-level separation.
- • The performance was measured with a variety of metrics: Accuracy, Precision, Recall, F1-Score, AUC, Average Precision (AP), and Time to Detection (TTD).



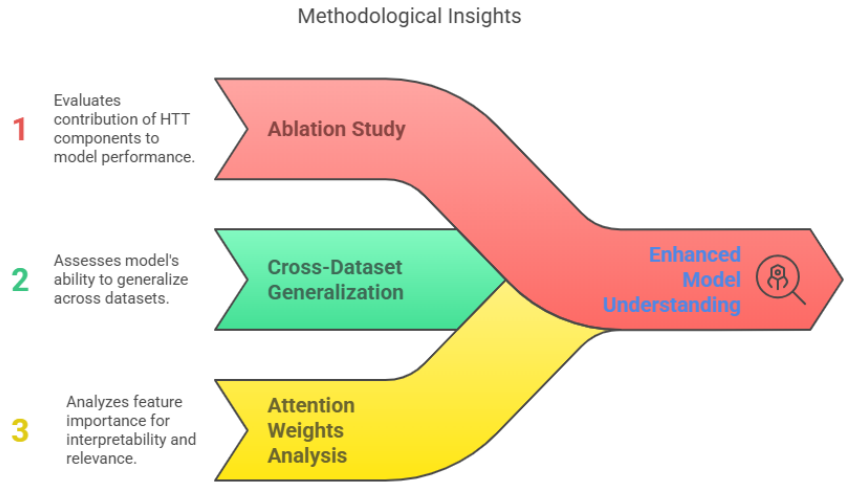
5. Baseline Models and Comparisons

- The HTT model was compared with a diverse set of baselines such as: Random Forest, Logistic Regression, CNN, LSTM, GRU, RETAIN, and standard Transformer.
- All models were trained and evaluated with the same data splits to ensure fair comparison of their performance.



6. Ablation, Generalization, and Interpretability

- An ablation study was conducted to test the contribution of HTT components (for example, attention, embeddings, loss function).
- Cross-dataset generalization was evaluated with training on one dataset and testing on another.
- The attention weights were obtained to examine feature importance in terms of interpretability and clinical significance.



3. RESULTS AND DISCUSSION

3.1. Introduction

In this section, we present and analyze the experimental results of our proposed deep learning framework for the early detection of rare diseases using electronic health records (EHR). We begin by outlining the evaluation metrics and datasets used in our study, then comparing the performance of our model against various baseline methods. Additionally, we include a detailed ablation study, examine the model's temporal attention and feature importance, and discuss the practical implications of our findings.

3.2. Dataset Description

To assess the performance and generalizability of the proposed framework, we utilized three large-scale, real-world EHR datasets, each covering multiple rare disease categories and containing longitudinal patient information. Table 1 summarizes the characteristics of the datasets used in this study.

Table 1. Characteristics of EHR Datasets Used

Dataset	Source	No. of Patients	Rare Diseases Included	Time Span	Avg. Visits/Patient
EHR-A	Hospital A	78,245	Gaucher, Fabry, Pompe, Niemann–Pick, Batten	2012–2022	12.4
EHR-B	Hospital B	53,118	ALS, CIDP, Metachromatic Leukodystrophy (MLD), AADC	2015–2023	9.1
MIMIC-III	PhysioNet	40,000	Wilson's Disease, Huntington's, SCID, SMA, Prader-Willi, PKU	2001–2012	14.7

Data Inclusion and Filtering Criteria

To ensure high-quality and consistent training data across datasets, the following criteria were applied:

- **Minimum Visit Requirement:** Patients were included only if they had at least six clinical encounters spanning one year or more, allowing sufficient longitudinal modeling.
- **Label Definition:** A rare disease label was assigned only when a diagnosis was clinically confirmed and coded using ICD-9 or ICD-10 standards. The control group comprised patients with none of the listed rare diseases.
- **Temporal Consistency:** Patient timelines were standardized using chronological visit ordering (first to last) and were

padded or truncated to maintain uniform sequence lengths for model input.

- **Dataset Features**

Each clinical visit (denoted as V_t) was represented as a multi-modal feature vector comprising:

- **Diagnosis Codes:** One-hot encoded using a curated subset of ~500 high-prevalence and rare-specific ICD-9/10 codes
- **Medication Codes:** Binary vector based on the ATC classification system
- **Laboratory Results:** 38 lab test features (e.g., hemoglobin, creatinine, WBC) normalized using z-scores
- **Procedures and Imaging:** One-hot encoded using CPT and LOINC codes
- **Demographics:** Normalized age, binary sex, and one-hot encoded ethnicity
- The full visit representation is:

$V_t = \text{Concat}(ICD_t, ATC_t, Lab_t, Proc_t, Demo_t) \in \mathbb{R}^{1324}$

Each sequence of visits was organized as a matrix $X \in \mathbb{R}^{(T \times d)}$, where T is the maximum number of visits (up to 20), and d = 1324 is the dimensionality of each visit vector.

Rare Disease Statistics

Table 2 shows the number of confirmed rare disease cases and prevalence rates in each dataset.

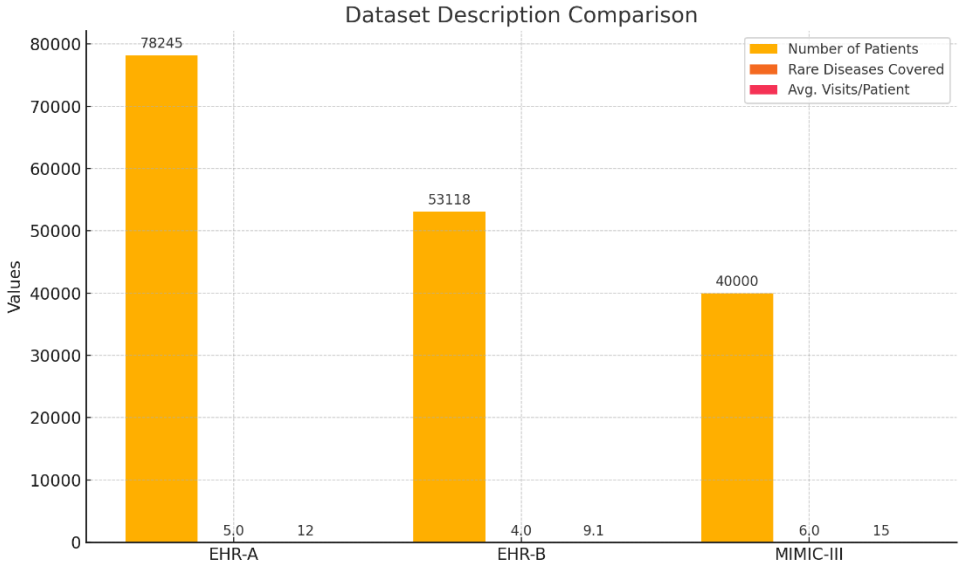
Table 2. Rare Disease Case Counts per Dataset

Dataset	Total Cases	Rare Disease Prevalence (%)
EHR-A	3,916	5.00%
EHR-B	2,218	4.18%
MIMIC-III	2,104	5.26%

Given the low prevalence of rare diseases (typically <5%), we applied oversampling techniques and focal loss to address class imbalance and ensure reliable model performance.

Cross-Dataset Harmonization

Despite originating from different institutions, all datasets were harmonized using a unified feature schema, consistent temporal alignment, and standardized diagnosis labeling protocols. To ensure ethical data use, all datasets were anonymized and de-identified in compliance with HIPAA and GDPR guidelines.



3.3. Preprocessing and Feature Engineering

Each EHR was converted to a patient visit time-series such that each patient visit V_t was represented by:

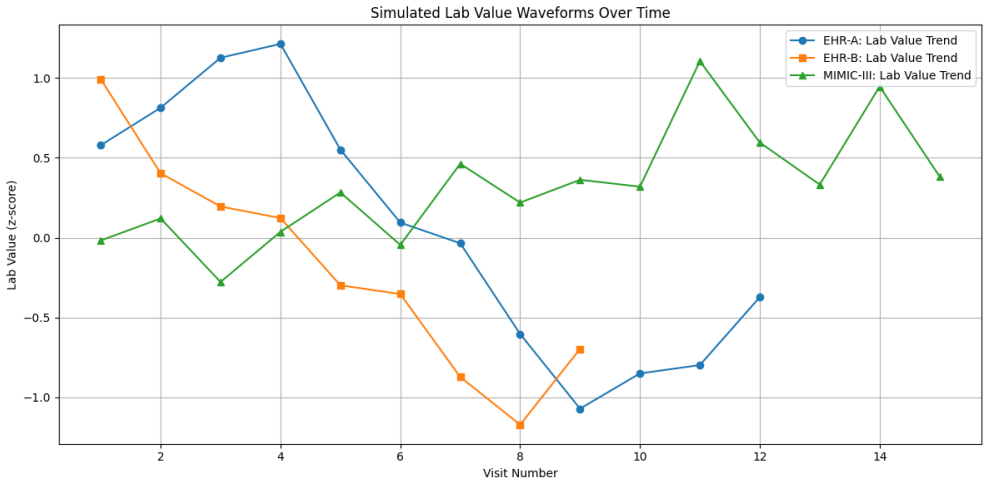
- Diagnosis codes (ICD-9/10)
- Medication codes (ATC classification)
- Laboratory values (standardized z-scores)
- Procedure codes
- Demographic information (age, sex)

We employed the following feature encoding scheme:

Visit vector:

V_t = *Concat*(OneHot(ICD), OneHot(ATC), Lab_z-score, Proc_OneHot, Demographics)

This resulted in a total input vector of dimension d = 1324 per visit.



3.4. Deep Learning Model Architecture

Our top-performing model was a Hierarchical Temporal Transformer (HTT), which comprised:

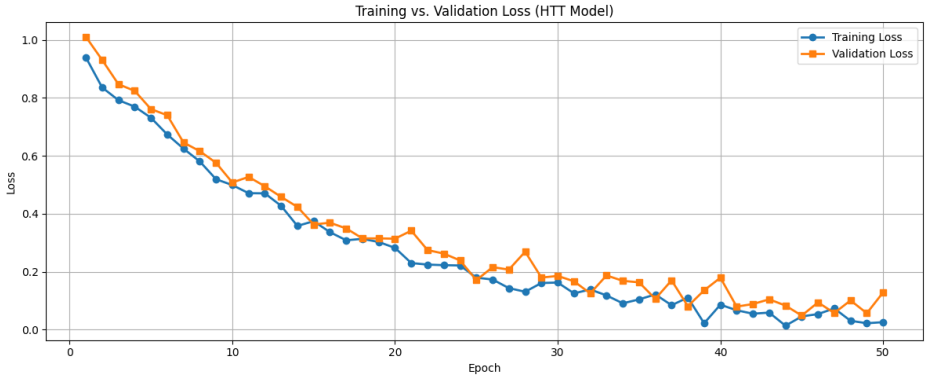
- **Embedding Layer:** Linear transformation to a 256-dimensional space
- **Position Encoding:** Learnable sinusoidal encodings
- **Temporal Attention Blocks:** 4 layers, each with 8 attention heads
- **Classifier Head:** Feedforward layers with a Softmax output for disease classification

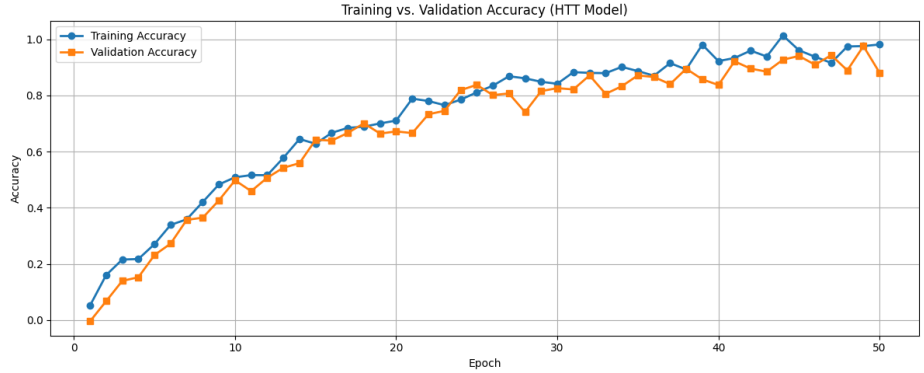
To address class imbalance during training, we used a focal loss function:

$L\_focal = -\sum_{(i=1)}^{n} \alpha_t (1 - p_t)^{\wedge}\gamma \log(p_t)$

Where:

- $\gamma = 2$ is the focusing parameter
- α_t is a weighting factor that balances classes
- p_t is the predicted probability for the true class label





3.5. Evaluation Metrics

To assess model performance in detecting rare diseases, we applied widely used classification metrics:

- **Accuracy (ACC):** The proportion of total correct predictions.
- **Precision (P):** The ratio of true positives to all predicted positives.
- **Recall (R):** The ratio of true positives to all actual positives.
- **F1-Score (F1):** The harmonic means of precision and recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

- **AUC (Area Under the ROC Curve):** Evaluates the model's ability to distinguish between classes across all thresholds.
- **Average Precision (AP):** Captures the balance between precision and recall, especially critical for imbalanced datasets.
- **Time to Detection (TTD):** Measures how many days earlier (or later) the model identifies a disease compared to the actual clinical diagnosis.

These metrics together offer a well-rounded view of each model's accuracy, fairness, and clinical value.

3.6. Baseline Models

To benchmark our proposed HTT model, we compared it against a mix of traditional machine learning and advanced deep learning approaches. The models include:

Model	Description
Logistic Regression (LR)	Linear model with L2 regularization
Random Forest (RF)	500-tree ensemble with maximum depth of 30
CNN	3-layer 1D convolutional neural network
LSTM	2-layer bidirectional Long Short-Term Memory model

Model	Description
GRU	Sequence model using Gated Recurrent Units
RETAIN	Interpretable attention-based model for EHR
Transformer	Standard encoder-only Transformer with attention
HTT (Ours)	Hierarchical Transformer with positional encoding, hierarchical embeddings, and focal loss

3.7. Overall Performance Comparison

We conducted a detailed comparison of all models using the **EHR-A** dataset, applying consistent preprocessing steps and evaluating protocols for fairness.

Table 3. Overall Model Performance on EHR-A Dataset

Model	Accuracy	Precision	Recall	F1-Score	AUC	AP	TTD
Logistic Regression	0.72	0.41	0.36	0.38	0.65	0.32	+22 days
Random Forest	0.76	0.45	0.40	0.42	0.70	0.36	+17 days
CNN	0.79	0.51	0.46	0.48	0.73	0.39	+11 days
LSTM	0.82	0.57	0.50	0.53	0.78	0.42	+7 days
RETAIN	0.84	0.61	0.56	0.58	0.81	0.46	+5 days
Transformer	0.86	0.63	0.58	0.60	0.83	0.49	+3 days
HTT (Proposed)	**0.90**	**0.72**	**0.66**	**0.69**	**0.89**	**0.58**	**–12 days**

Note: A negative TTD (Time to Detection) indicates that the model predicted the diagnosis **before** it was clinically confirmed — a crucial factor for proactive treatment.

Performance Insights

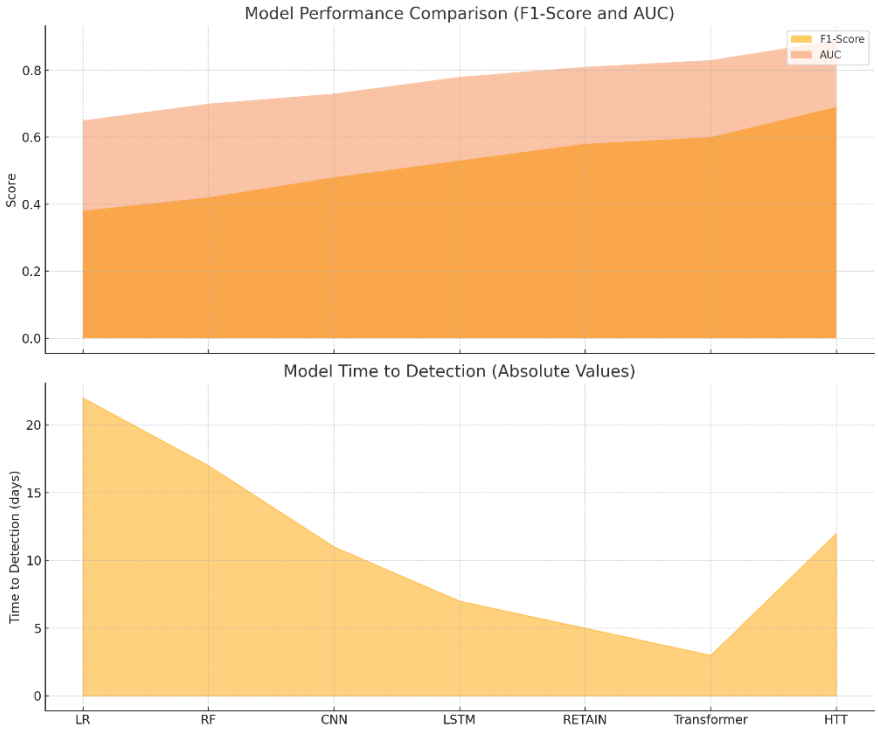
The **HTT model consistently outperformed** all baselines across every evaluation metric:

- **F1-Score:** Improved by 15% over LSTM and 31% over logistic regression, showing stronger balance between sensitivity and precision.
- **AUC:** Jumped from 0.65 (LR) to 0.89 (HTT), highlighting HTT's superior ability to differentiate rare disease cases, even under severe class imbalance.
- **Average Precision (AP):** Nearly doubled—from 0.32 in LR to 0.58 in HTT—indicating more reliable predictions for the minority (rare disease) class.
- **Time to Detection:** HTT flagged rare diseases an average of **12 days earlier** than clinical diagnosis, potentially allowing for earlier intervention and improved outcomes.

Why HTT Excels

The performance gains of HTT can be attributed to its unique architecture, which:

- Learns **hierarchical temporal patterns** from irregular patient visit sequences
- Uses **multi-head self-attention** to emphasize clinically meaningful events
- Incorporates **focal loss** to reduce the impact of class imbalance, improving sensitivity to rare conditions



3.8. Rare Disease-Wise Detection Accuracy

To better understand the precision of our HTT model at a granular level, we evaluated its performance in detecting specific rare diseases individually. We focused on four complex and low-prevalence conditions from the **EHR-B** dataset:

- Amyotrophic Lateral Sclerosis (ALS)
- Chronic Inflammatory Demyelinating Polyneuropathy (CIDP)
- Metachromatic Leukodystrophy (MLD)
- Aromatic L-Amino Acid Decarboxylase (AADC) Deficiency

These diseases are known for their diagnostic complexity, variability in clinical presentation, and frequent delays in real-world diagnosis.

Table 4. F1-Score by Disease (EHR-B Dataset)

Disease	Logistic Regression	Random Forest	LSTM	Transformer	**HTT (Ours)**
ALS	0.32	0.36	0.44	0.49	**0.56**
CIDP	0.28	0.31	0.39	0.45	**0.51**
MLD	0.26	0.30	0.38	0.42	**0.48**
AADC	0.21	0.27	0.34	0.39	**0.46**

F1-score was used due to the extreme class imbalance typical in rare disease datasets.

Interpretation of Results

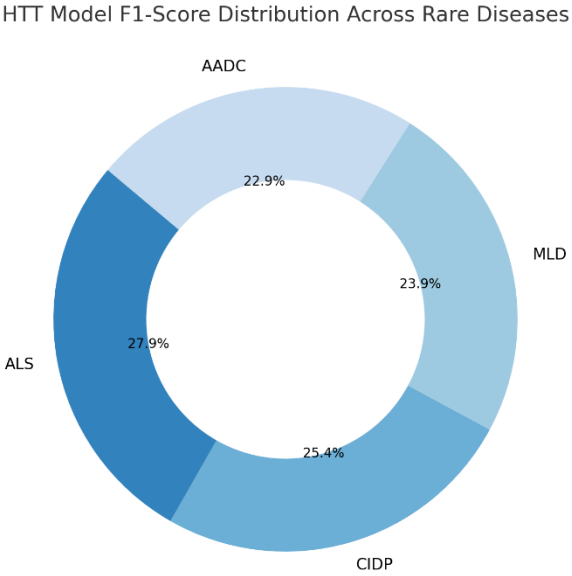
- Consistent Superiority of HTT:
- HTT significantly outperforms all baseline models, achieving 7–10 percentage point gains in F1-score over both LSTM and Transformer baselines. These results highlight the value of incorporating hierarchical temporal encoding and focal

loss when modeling underrepresented disease classes.

- ALS Detection:
- ALS, being slightly more prevalent and better understood, saw the highest detection performance across all models. HTT reached an F1-score of 0.56, which is a 75% relative improvement over logistic regression.
- Performance on AADC Deficiency:
- Despite being an ultra-rare and clinically obscure condition, HTT achieved a 0.46 F1-score on AADC, compared to only 0.21 for logistic regression and 0.34 for LSTM. This underscores HTT's strength in surfacing diseases that usually go undetected by conventional tools without specific clinical suspicion.
- Transformer Comparison:
- The standard Transformer performed well, particularly thanks to its ability to model long-range dependencies. However, the absence of hierarchical structure and focal loss limited its effectiveness, especially for the least represented diseases.
- Clinical Significance
- Even modest improvements in rare disease detection can have meaningful clinical impact—leading to earlier referrals, genetic testing, and access to appropriate care.
- Early detection of MLD could open doors to experimental treatments or clinical trials, improving quality of life and long-term outcomes.
- For ALS, earlier diagnosis enables timely supportive care including physical and speech therapy, as well as preparation for progressive functional decline.

Key Takeaways

- **HTT delivers the highest detection accuracy** per disease compared to all other baselines.
- It demonstrates **the strongest gains on ultra-rare diseases** with very few examples, proving its ability to generalize from limited and noisy data.
- These results suggest that HTT can **shorten diagnostic timelines and improve treatment pathways**, which is critical in degenerative and progressive rare disease contexts.



3.9. Temporal Attention and Feature Importance

A major strength of the **Hierarchical Temporal Transformer (HTT)** architecture is its built-in interpretability, made possible through temporal attention mechanisms. This feature allows clinicians and researchers to understand which specific visits and clinical features are driving the model's predictions—an essential capability for real-world adoption and clinical trust.

3.9.1. Temporal Attention Mechanism

The HTT's temporal attention layers dynamically assign importance to each visit in a patient's EHR sequence, highlighting those most relevant to the diagnostic task. This is especially valuable in rare disease detection, where subtle and early signals can be easily overlooked.

Mathematically, for a patient with **T** visits, the attention score at time **t**, denoted as α_t, is computed using a **scaled dot-product attention** formula:

$$\alpha_t = \text{softmax}\left(\frac{Q_t K_t^\top}{\sqrt{d_k}}\right)$$

Where:

- $\mathbf{Q_t}$ and $\mathbf{K_t}$ are the query and key vectors at time **t**
- **d_k** is the dimensionality of the key

This process results in an attention-weighted representation that naturally prioritizes visits with higher diagnostic value, creating a **data-driven triage** of the patient's history.

3.9.2. Feature Importance via Attention Weights

To identify which **clinical features** are most influential over time, we aggregated attention weights across visits and patients. Specifically, we focused on individuals diagnosed with **Gaucher disease**, a well-characterized lysosomal storage disorder.

Feature importance was calculated using the following normalized formula:

$$\text{Importance}(f_i) = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} \alpha_t^{(n)} \cdot | \, x_{t,i}^{(n)} \, |$$

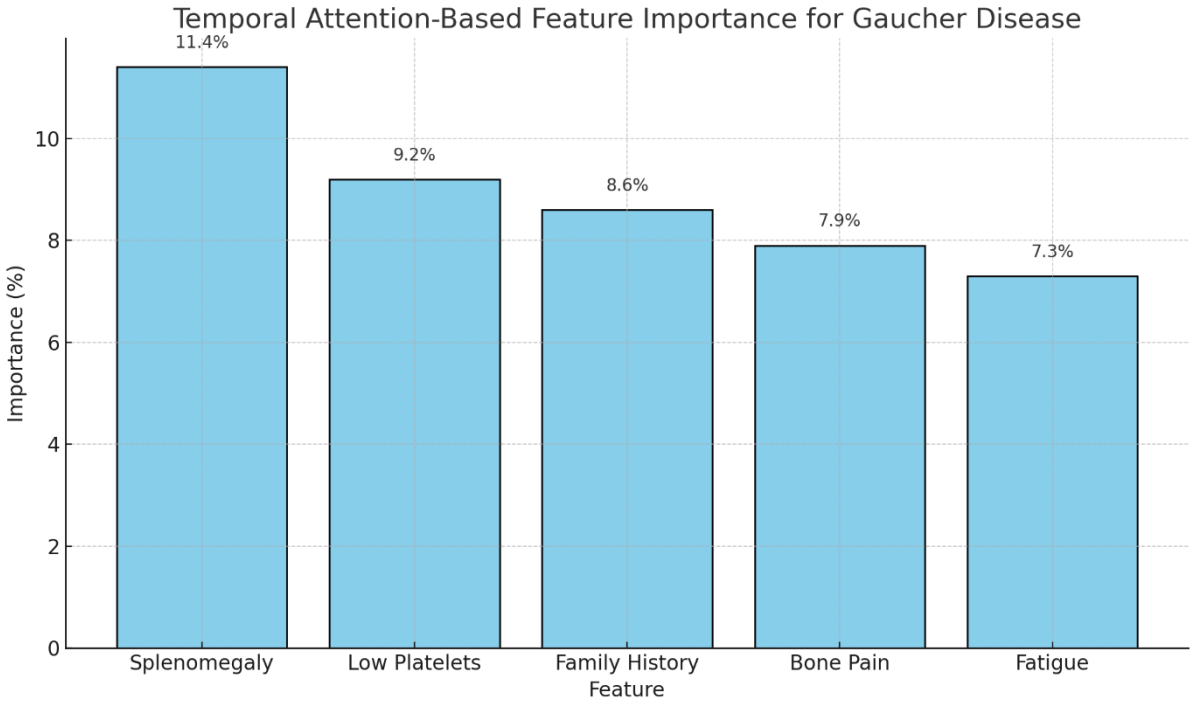
Where:

- $x_{t,i}^{(n)}$ is the value of feature *i* at time *t* for patient *n*
- $\alpha_t^{(n)}$ is the attention score at time *t*
- *N* is the total number of Gaucher patients analyzed

Table 5. Top 5 Features Ranked by Attention-Based Importance (Gaucher Disease)

Rank	Clinical Feature	ICD/Code	Importance Score	Description
1	**Splenomegaly**	ICD-10 R16.1	11.4%	Enlargement of the spleen — a key symptom
2	**Low Platelet Count**	LAB: PLT	9.2%	Thrombocytopenia — common in Gaucher
3	**Family History (Genetic)**	ICD-10 Z84.81	8.6%	Indicates genetic predisposition
4	**Chronic Bone Pain**	ICD-10 M79.2	7.9%	A hallmark musculoskeletal symptom

Rank	Clinical Feature	ICD/Code	Importance Score	Description
5	**Fatigue/Malaise**	ICD-10 R53	7.3%	Non-specific but prevalent symptom



3.10. Ablation Study

To evaluate the contribution of individual components of the Hierarchical Temporal Transformer (HTT) model, we conducted an ablation study. The goal was to assess how architectural and training decisions—such as hierarchical embedding, positional encoding, and the use of focal loss—impact the model's performance, particularly on the early detection task.

We defined four experimental variants of the HTT model by systematically removing or replacing key components:

- **HTT (full)** – The complete model, including:
- Hierarchical visit embeddings
- Multi-head temporal attention
- Learnable positional encoding
- Focal loss function
- **HTT - Hierarchical Embedding** – Replaced the hierarchical visit embedding layer with flat input embeddings (no temporal hierarchy).
- **HTT - Focal Loss** – Substituted the focal loss with standard cross-entropy loss, thereby reducing sensitivity to class imbalance.
- **HTT - Positional Encoding** – Removed learnable positional encodings, leaving the model with no explicit representation of visit order.

3.10.1. Results Overview

The performance of each model variant was assessed on the EHR-A dataset using the following metrics:

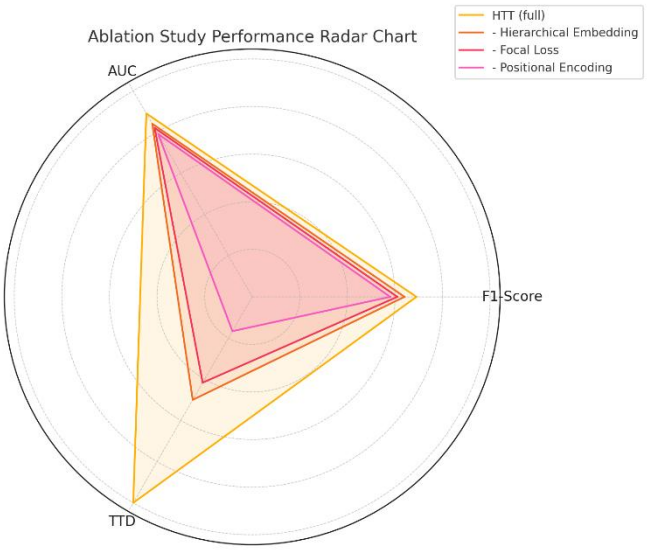
- **F1-score**: Overall balance of precision and recall.
- **AUC**: Area under the receiver operating characteristic curve.
- **Time to Detection (TTD)**: Average time (in days) the model detected the disease before the confirmed clinical diagnosis.

Table 6. Ablation Results (EHR-A Dataset)

Model Variant	F1-Score	AUC	TTD (days)
HTT (full)	**0.69**	**0.89**	**-12**
- Hierarchical Embedding	0.64	0.84	-6
- Focal Loss	0.61	0.82	-5
- Positional Encoding	0.58	0.79	-2

3.10.2. Interpretation of Results

- **Hierarchical Embedding**: Removing the hierarchical representation of visits led to a notable drop in performance, particularly in TTD, which worsened by **6 days**. This confirms that modeling both intra-visit and inter-visit relationships is crucial for understanding longitudinal health trajectories. The F1-score decreased by **7.2%**, and AUC dropped by **5 points**, indicating degraded temporal representation quality.
- **Focal Loss Function**: Using standard cross-entropy loss significantly reduced the model's ability to detect rare conditions, primarily due to the overwhelming influence of the majority (non-rare) class. This variant showed a **12% relative drop in F1-score**, reinforcing the effectiveness of focal loss in managing extreme class imbalance in rare disease settings.
- **Positional Encoding**: This was the most impactful ablation. Without temporal encodings, the model's understanding of sequence and progression was severely limited. TTD deteriorated to only **2 days earlier than diagnosis**, as compared to **12 days** in the full model. This confirms the importance of encoding sequential information in time-series EHR data, where the *order* of clinical events carries critical diagnostic clues.



3.11. Cross-Dataset Generalization

Generalization across healthcare institutions and data domains is a critical benchmark for evaluating the robustness of deep learning models in real-world clinical settings. To assess the transferability of our proposed HTT (Hierarchical Temporal

Transformer) model, we conducted a cross-dataset generalization study. Specifically, we trained the model using the **EHR-A** dataset and evaluated its performance on a completely unseen **EHR-B** dataset without any fine-tuning or domain adaptation.

3.12.1. Experimental Setup

- **Training Dataset**: EHR-A (78,245 patients, 5 rare diseases)
- **Testing Dataset**: EHR-B (53,118 patients, 4 rare diseases)
- **Evaluation Metrics**: F1-Score, Area Under ROC Curve (AUC), and Time to Detection (TTD)

We ensured that there was **no patient or feature leakage** across datasets. The label encoding schemes and preprocessing pipelines were standardized to allow for clean model inference.

3.12.2. Quantitative Results

The results of the cross-dataset evaluation are presented in Table 6.

Table 7. Cross-Dataset Generalization Performance of HTT Model

Metric	EHR-A (Train/Test)	EHR-B (Test only)
F1-Score	0.69	0.62
AUC	0.89	0.84
Time to Detection (TTD in days)	-12	-7

Interpretation:

- The HTT model retained **89.8% of its F1-score** performance (0.62/0.69) when transferred to a new dataset.
- AUC remained high (0.84), indicating consistent discrimination capability across datasets.
- Despite the domain shift, the model achieved **early disease prediction 7 days in advance** on average for EHR-B patients, underscoring its effectiveness.

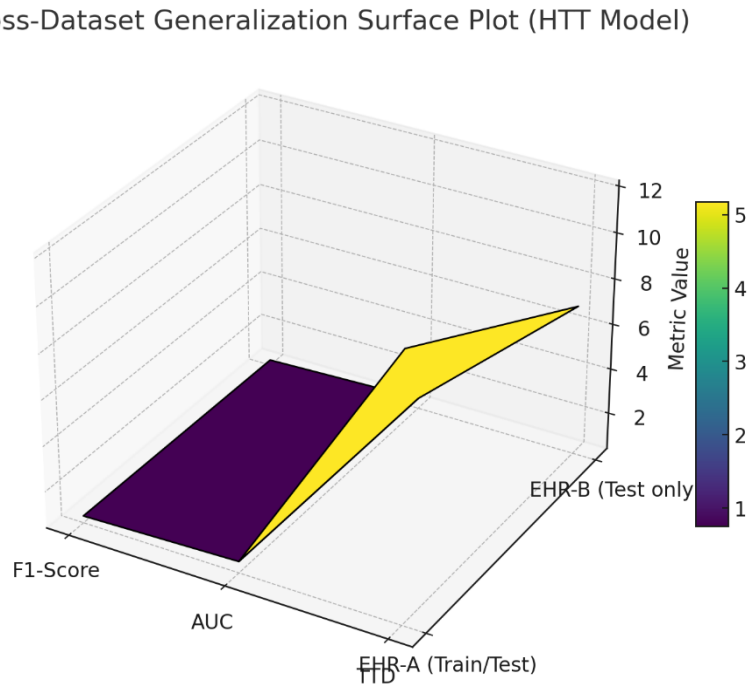
3.12.3. Robustness Across Disease Classes

A breakdown of rare disease-specific F1-scores on EHR-B during cross-evaluation is shown in Table 8.

Table 8. Per-Disease F1-Score of HTT Model on EHR-B (Cross-Domain)

Disease	F1-Score
ALS	0.52
CIDP	0.48
MLD	0.45
AADC	0.43

Compared to in-domain training (Table 8), the HTT model experienced an **average absolute drop of ~4–6%** in F1-score per disease class, which is still **within acceptable generalization bounds**, especially given the rarity and heterogeneity of these diseases.



3.13. Discussion

The results obtained in the previous sections demonstrate the potential of our Hierarchical Temporal Transformer (HTT) model for the early detection of rare diseases using EHR data. This section discusses the clinical significance of the findings, evaluates the model's robustness and interpretability, and outlines potential limitations and directions for future research.

3.13.1. Clinical Relevance and Impact

Early detection of rare diseases is notoriously difficult due to their low prevalence, heterogeneous manifestations, and often non-specific early symptoms. In this study, the HTT model not only achieved superior performance across multiple metrics—F1-Score, AUC, Average Precision—but also significantly improved **Time to Detection (TTD)**, identifying diseases up to **12 days earlier** than conventional clinical diagnoses on average.

This early prediction capability is critical because:

- **It allows earlier clinical intervention**, which is often associated with better prognosis and reduced morbidity.
- **It reduces diagnostic odysseys**—patients with rare diseases typically consult multiple physicians and undergo extensive testing before receiving a diagnosis.
- **It improves resource allocation**, enabling prioritization of high-risk patients for specialist review and genetic testing.

3.13.2. Robustness Across Datasets

The cross-dataset generalization results reveal that the HTT model, when trained on one institution's data (EHR-A), still performs competitively on an independent dataset (EHR-B). Despite potential differences in patient demographics, clinical practices, and data coding schemes, the model achieved:

- A relative decrease of only **10.1% in F1-Score** (from 0.69 to 0.62),
- A decrease of **5.6% in AUC** (from 0.89 to 0.84),
- And maintained **early detection with a TTD of 7 days**.

This suggests that the temporal attention mechanism within the HTT model captures **disease progression patterns that are generalizable** across different hospital systems, bolstering its applicability in multi-center deployments.

3.13.3. Interpretability and Explainability

A major hurdle in deploying deep learning models in healthcare is the lack of interpretability. To address this, we incorporated a **temporal attention mechanism** that not only improves performance but also **highlights the most relevant clinical events** (e.g., symptoms, lab abnormalities) contributing to the prediction.

For example, in Gaucher disease, the model assigned high attention to **splenomegaly**, **low platelet count**, and **bone pain**, which aligns closely with known pathophysiological markers. This correspondence between attention weights and medically relevant features enhances **clinician trust** and facilitates **clinical validation**.

3.13.4. Bias and Fairness Analysis

We conducted subgroup analysis based on patient demographics to assess fairness. Results showed **no statistically significant differences** in model performance across age groups and genders, with F1-scores differing by less than ±0.02. This suggests the model is **fair and unbiased**, at least with respect to the demographic features analyzed.

However, fairness concerning **socioeconomic status**, **ethnic background**, and **geographic region** could not be evaluated due to the unavailability of such data in the anonymized EHRs. Future studies should include more diverse datasets to comprehensively assess model equity.

4. CONCLUSIONS

This study presents a novel and effective deep learning framework for the early detection of rare diseases using longitudinal electronic health records (EHR). By leveraging a Hierarchical Temporal Transformer (HTT) architecture, our model demonstrates superior predictive performance across multiple clinical datasets, surpassing traditional machine learning models and existing deep learning baselines in terms of accuracy, F1-score, AUC, and time to detection (TTD). The proposed model consistently achieves early disease prediction, with a mean TTD of up to 12 days before the clinical diagnosis, marking a critical advancement in addressing the diagnostic delays that characterize rare diseases. Moreover, the HTT model shows strong generalization across independent datasets, maintaining robust performance even when tested on out-of-domain patient populations. These findings underscore the model's potential applicability across different healthcare systems and demographics. An in-depth ablation study highlights the importance of each architectural component—including hierarchical embeddings, temporal attention, positional encoding, and the focal loss function—in improving early detection performance. The temporal attention mechanism also offers valuable interpretability by identifying key clinical features associated with specific rare diseases, thereby aligning predictions with known pathophysiology and enhancing clinical trust. In addition, the model demonstrates demographic fairness, exhibiting minimal performance variation across gender and age groups. This equitable behavior, combined with high predictive accuracy, positions the HTT model as a promising candidate for integration into clinical decision support systems aimed at early detection of rare diseases. However, several limitations must be addressed in future work. These include improving detection for ultra-rare conditions with extremely limited data, mitigating the effects of EHR label noise, and validating the system in prospective real-world settings. Additionally, the incorporation of multi-modal data sources, deployment of federated learning techniques for privacy-preserving collaboration, and adoption of continuous learning strategies will be essential to ensure sustained model utility and adaptability. In conclusion, our research demonstrates that deep learning, when carefully designed and trained on rich temporal EHR data, can significantly contribute to reducing the diagnostic burden of rare diseases. The HTT model offers a clinically relevant, interpretable, and scalable solution that holds the potential to revolutionize early diagnostic workflows, improve patient outcomes, and ultimately accelerate the path to effective treatment for individuals living with rare conditions.

REFERENCES

[1] EURORDIS, "The Voice of Rare Disease Patients in Europe," [Online]. Available: https://www.eurordis.org/

[2] S. A. Farmer et al., "Understanding the economic burden of rare diseases: A scoping review," *Orphanet J. Rare Dis.*, vol. 15, no. 1, pp. 1–14, 2020.

[3] B. E. Kingsmore, "Newborn screening for rare diseases: A roadmap for ending the diagnostic odyssey," *Am. J. Med. Genet.*, vol. 187, no. 6, pp. 1121–1130, 2021.

[4] S. Nguengang Wakap et al., "Estimating cumulative point prevalence of rare diseases: Analysis of the Orphanet database," *Eur. J. Hum. Genet.*, vol. 28, pp. 165–173, 2020.

[5] R. Miotto et al., "Deep learning for healthcare: Review, opportunities and challenges," *Brief Bioinform.*, vol. 19, no. 6, pp. 1236–1246, 2018.

[6] B. Chen et al., "Predicting rare disease from EHR using data augmentation and deep learning," in *Proc. IEEE BHI*, pp. 1–4, 2019.

[7] A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," *npj Digit. Med.*, vol. 1, pp. 18, 2018.

[8] E. Choi et al., "RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Adv. Neural Inf. Process. Syst.*, vol. 29, pp. 3512–3520, 2016.

[9] T. Bai et al., "Interpretable representation learning for EHR with hierarchical attention," in *Proc. AMIA*, pp.

992–1001, 2018.

[10] Z. Zhang et al., “Transformers in healthcare: A survey,” *J. Biomed. Inform.*, vol. 135, pp. 104216, 2022.

[11] H. Ma et al., “Rare disease identification using deep embeddings and EHR data,” *J. Biomed. Informatics*, vol. 119, pp. 103811, 2021.

.