https://www.scirp.org/journal/jbise

J. Biomedical Science and Engineering,
2026, Vol. 19, (No. 1), pp: 8-14

Symptom Cascade Analyzer: A Graph-Theoretic Natural Language Processing Framework for Culturally-Adaptive Medical Diagnosis

Felix Davis

Department of Computer Science, Dartmouth College, Hanover, NH, USA

Correspondence to: Felix Davis, felix.d.davis.26@dartmouth.edu

Keywords: Multilingual, Natural Language Processing, Medical Diagnosis, Knowledge Graphs, Cascade Analysis
Received: August 3, 2025 **Accepted:** January 9, 2026 **Published:** January 12, 2026

ABSTRACT

We present the Symptom Cascade Analyzer (SCA), a natural language processing framework for culturally-adaptive medical diagnosis that integrates graph-theoretic symptom modeling, multilingual embeddings, and cultural adaptation layers. The framework incorporates graph entropy for rare-disease detection and demonstrates a 23% improvement in diagnostic accuracy for culturally specific symptom descriptions. Spectral clustering entropy analysis further enhances the identification of rare diseases. These results highlight SCA's potential for deployment in multilingual, culturally diverse clinical environments.

1. INTRODUCTION

1.1. Problem Formulation

Medical diagnosis from natural language symptom descriptions presents a complex multi-objective optimization problem [1, 2]. Let $\mathcal{S}=\{s_1,s_2,\cdots,s_n\}$ denote the symptom space and $\mathcal{D}=\{d_1,d_2,\cdots,d_m\}$ represent the disease space. The diagnostic mapping seeks to construct $\mathcal{F}:\mathcal{L}\to\mathcal{P}(\mathcal{D})$ where $\mathcal{L}$ is the linguistic input space and $\mathcal{P}(\mathcal{D})$ represents probability distributions over diseases.

The cultural adaptation challenge requires modeling the function:

$$\Gamma:\mathcal{L}\times\mathcal{C}\to\mathbb{R}^d \tag{1}$$

where $\mathcal{C}$ represents cultural contexts and d is the embedding dimension.

The optimization objective incorporates both diagnostic accuracy and cultural sensitivity:

$$\theta^*=\arg\min_\theta \mathbb{E}_{(\ell,c,d)\sim P}\left[L\left(F_\theta\left(\Gamma\left(\ell,c\right)\right),d\right)\right]+\lambda_1 R(\theta)+\lambda_2 H(G). \tag{2}$$

This modification makes explicit the entropy regularization term $H(G)$ referenced in the abstract. Where:

- $\mathcal{P}$, is defined as the empirical data distribution in the triples of symptoms-culture-disease $((\ell, c, d))$ induced by the training corpus. Expectations are approximated using minibatches sampled uniformly from this empirical dataset.
- L is also defined as the cross-entropy loss between the predicted disease distribution $\mathcal{L}\left(\mathcal{F}_\theta\left(\Gamma(\ell, c)\right)\right)$ and the ground-truth label $\mathbb{E}_d$.

1.2. Graph-Theoretic Foundation

The SCA framework models medical knowledge as a directed weighted graph [3]. That is $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ where:

$$\mathcal{V} = \mathcal{S} \cup \mathcal{D} \quad (\text{symptoms and diseases}) \tag{3}$$

$$\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V} \quad (\text{causal relationships}) \tag{4}$$

$$\mathcal{W} : \mathcal{E} \to \mathbb{R}^+ \quad (\text{probabilistic weights}) \tag{5}$$

The edge weights $w_{i,j} \in \mathcal{W}$ represent conditional probabilities $P(v_j \mid v_i)$ where $v_i, v_j \in \mathcal{V}$. These weights are adjusted by cultural specificity factors:

$$w_{i,j}^{(c)} = w_{i,j} \cdot \text{Gloss}(c, v_i, v_j) \tag{6}$$

where $\text{Gloss}(c, v_i, v_j)$ represents the cultural glossary mapping function.

2. METHODS

2.1. Cultural Glossary Integration

The cultural adaptation mechanism employs a hierarchical glossary structure [4], $\mathcal{G}_c = \{\mathcal{T}_c, \mathcal{M}_c, \mathcal{E}_c\}$ where:

$$\mathcal{T}_c : \text{cultural terminology mappings} \tag{7}$$

$$\mathcal{M}_c : \text{metaphorical expression translations} \tag{8}$$

$$\mathcal{E}_c : \text{ethnic bias adjustment factors} \tag{9}$$

The glossary mapping function is defined as:

$$\text{Gloss}(c, \ell) = \sum_{t \in \mathcal{T}_c} \alpha_t \phi_t(\ell) + \sum_{m \in \mathcal{M}_c} \beta_m \psi_m(\ell) + \gamma_c \xi_c(\ell). \tag{10}$$

The term $\gamma_c \xi_c(\ell)$ in Equation (10) corresponds to the ethnic bias adjustment component E_c, where γ_c is the learned culture-specific weight and $\xi_c(\ell)$ extracts ethnicity-dependent linguistic features. Where ϕ_t, ψ_m, ξ_c are feature extraction functions and $\alpha_t, \beta_m, \gamma_c$ are learned weights.

For example, the cultural mapping "stomach fire" $\to$ "gastritis" is encoded [5] as:

$$\mathcal{T}_c(\text{"stomachfire"}) = \arg\max_{d \in \mathcal{D}} P(d \mid \text{"stomachfire"}, c) \cdot \text{sim}(\text{"stomachfire"}, d) \tag{11}$$

2.2. Multilingual Embedding Architecture

The SCA employs XLM-RoBERTa for multilingual representation learning. Let $\mathbf{h}_\ell \in \mathbb{R}^d$ denote the embedding for linguistic input ℓ. The cultural adaptation layer transforms embeddings via:

$$\mathbf{h}_\ell^{(c)} = \mathbf{W}_c \mathbf{h}_\ell + \mathbf{b}_c + \mathbf{A}_c \odot \mathbf{h}_\ell \tag{12}$$

where $\mathbf{W}_c \in \mathbb{R}^{d \times d}$ is the cultural transformation matrix, $\mathbf{b}_c \in \mathbb{R}^d$ is the cultural bias vector, $\mathbf{A}_c \in \mathbb{R}^d$ provides element-wise scaling, and $\odot$ denotes Hadamard product.

The attention mechanism for cultural context is:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathbf{M}_c\right)\mathbf{V} \tag{13}$$

where $\mathbf{M}_c$ is the cultural attention mask.

2.3. Bias Adjustment Mechanisms

The SCA incorporates a parameterized bias-adjustment mechanism inspired by clinical severity-duration scoring models, as formalized in (14). Let $s_{i,j}$ represent symptom severity and $\tau_{i,j}$ denote duration. The adjusted edge weight is:

$$w_{i,j}^{\text{adj}} = w_{i,j} \cdot \exp\left(\alpha_s s_{i,j} + \alpha_\tau \log\left(\tau_{i,j} + 1\right)\right) \tag{14}$$

where α_s, α_τ are learned scaling parameters.

The bias adjustment function incorporates demographic factors:

$$\mathcal{B}(\mathbf{d}, c) = \prod_{k=1}^{K} \left(1 + \beta_k \mathbf{d}_k\right)^{\gamma_{c,k}} \tag{15}$$

where $\mathbf{d} \in \mathbb{R}^K$ represents demographic features and $\gamma_{c,k}$ are culture-specific exponents.

2.4. Rare Disease Detection via Cluster Entropy

Rare disease detection [6] employs cluster entropy analysis [7] on the symptom graph. Let $\mathcal{C} = \{C_1, C_2, \cdots, C_K\}$ denote symptom clusters obtained via spectral clustering on the adjacency matrix $\mathbf{A} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$.

The cluster entropy for cluster C_k is:

$$H(C_k) = -\sum_{v_i \in C_k} P(v_i \mid C_k) \log P(v_i \mid C_k) \tag{16}$$

Rare diseases are identified by low-entropy clusters:

$$\mathcal{D}_{\text{rare}} = \left\{ d \in \mathcal{D} : d \in C_k \text{ and } H(C_k) < \tau_{\text{entropy}} \right\} \tag{17}$$

The rare disease probability is computed as:

$$P(d \in \mathcal{D}_{\text{rare}} \mid \mathbf{s}) = \sigma\left(\mathbf{w}_{\text{rare}}^{\text{T}} \mathbf{f}(\mathbf{s}) - \log H(C_d)\right) \tag{18}$$

where $\mathbf{f}(\mathbf{s})$ extracts features from symptom vector $\mathbf{s}$ and σ is the sigmoid function, $\mathbf{f}(\mathbf{s})$ can be expressed as:

$$f(s) = \left[\text{freq}(s), h_s^{\text{GAT}}, \mathbf{1}(s \in C_k)\right] \tag{19}$$

Here, $\text{freq}(s)$ denotes symptom frequency counts, h_s^{GAT} represents the GAT-derived symptom embedding, and $\mathbf{1}(s \in C_k)$ indicates cluster membership for rare-disease detection.

2.5. Graph Neural Network Architecture

The SCA employs a Graph Attention Network (GAT) [8] for symptom-disease relationship modeling. The attention mechanism for node i with neighbors $\mathcal{N}(i)$ is:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^{\text{T}}\left[\mathbf{W}\mathbf{h}_i \,\|\, \mathbf{W}\mathbf{h}_j\right]\right)\right)}{\sum_{k \in \mathcal{N}(i)} \exp\left(\text{LeakyReLU}\left(\mathbf{a}^{\text{T}}\left[\mathbf{W}\mathbf{h}_i \,\|\, \mathbf{W}\mathbf{h}_k\right]\right)\right)} \tag{20}$$

The updated node representation is:

$$\mathbf{h}_i' = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}\mathbf{W}\mathbf{h}_j\right) \tag{21}$$

Multi-head attention aggregates information:

$$\mathbf{h}_i^{(M)} = \|_{m=1}^{M}\sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(m)}\mathbf{W}^{(m)}\mathbf{h}_j\right) \tag{22}$$

where $\|$ denotes concatenation and M is the number of attention heads.

3. ALGORITHM SPECIFICATION

Algorithm 1. **Symptom Cascade Analyzer (SCA).**

Data: Symptom description ℓ, cultural context c, graph $\mathcal{G}$
Result: Disease probability distribution $\mathbf{p}\in\mathbb{R}^{|\mathcal{D}|}$
$\mathbf{h}_\ell \leftarrow \text{XLM}-\text{RoBERTa}(\ell)$;
$\mathbf{h}_\ell^{(c)} \leftarrow \text{CulturalAdaptation}(\mathbf{h}_\ell, c)$;
$\mathbf{s} \leftarrow \text{ExtractSymptoms}(\mathbf{h}_\ell^{(c)})$;
for $(s_i, d_j)\in\mathcal{E}$;
$w_{i,j}^{(c)} \leftarrow w_{i,j}\cdot\text{Gloss}(c, s_i, d_j)$;
$w_{i,j}^{\text{adj}} \leftarrow w_{i,j}^{(c)}\cdot\mathcal{B}(\text{severity}, \text{duration})$;
end
$\mathbf{h}_{\text{nodes}} \leftarrow \text{GAT}(\mathcal{G}, \mathbf{h}_\ell^{(c)})$;
$\mathcal{C} \leftarrow \text{SpectralClustering}(\mathbf{A})$;
$H_{\text{clusters}} \leftarrow \text{ComputeClusterEntropy}(\mathcal{C})$;
$\mathbf{p}_{\text{common}} \leftarrow \text{softmax}(\mathbf{W}_{\text{out}}\mathbf{h}_{\text{diseases}})$;
$\mathbf{p}_{\text{rare}} \leftarrow \text{RareDiseaseDetection}(\mathbf{s}, H_{\text{clusters}})$;
$\mathbf{p} \leftarrow \lambda\mathbf{p}_{\text{common}} + (1-\lambda)\mathbf{p}_{\text{rare}}$;
return $\mathbf{p}$

4. RESULTS AND DISCUSSION

4.1. Experimental Setup

Evaluation was conducted on a multilingual medical corpus containing $N = 15000$ symptom-disease pairs across 12 languages and 8 cultural contexts. The dataset included both common and rare diseases with cultural-specific symptom descriptions.

4.2. Performance Metrics

The SCA algorithm achieved the following performance:

- Overall diagnostic accuracy: 87.3% ± 2.1%
- Cultural adaptation improvement: +23% over baseline
- Rare disease detection F1-score: 0.78 ± 0.05
- Multilingual consistency: $\kappa = 0.82$

The consistency of multilingual diagnostics was measured using Cohen's κ, treating each language-specific model as an independent rating. For each symptom description, the predicted disease categories were compared between languages and κ was calculated over these categorical ratings to quantify the agreement beyond chance.

4.3. Cultural Adaptation Analysis

Table 1 presents performance across cultural contexts:

Table 1. Performance analysis across cultural contexts.

Cultural Context	Accuracy (%)	Precision	Recall
Western/English	91.2 ± 1.8	0.89	0.93
Sub-Saharan Africa	85.7 ± 2.3	0.83	0.88
East Asian	88.4 ± 2.0	0.86	0.91
Latin American	86.9 ± 2.2	0.84	0.89
Middle Eastern	84.3 ± 2.5	0.81	0.87
South Asian	87.1 ± 2.1	0.85	0.90

4.4. Rare Disease Detection

The cluster entropy approach successfully identified rare diseases with:

- Sensitivity: 0.74 ± 0.06 for diseases with prevalence <1%
- Specificity: 0.92 ± 0.03 for common disease exclusion
- Positive predictive value: 0.68 ± 0.08

4.5. Ablation Study

Table 2 demonstrates component contributions:

Table 2. Ablation study for SCA components.

Configuration	Accuracy (%)	Cultural Improvement
Baseline NLP	64.2 ± 3.1	-
+ Multilingual Embeddings	71.8 ± 2.7	+7.6%
+ Cultural Glossary	79.3 ± 2.4	+15.1%
+ Graph Neural Network	83.7 ± 2.2	+19.5%
+ Rare Disease Detection	87.3 ± 2.1	+23.1%

5. CONCLUSION

The Symptom Cascade Analyzer demonstrates significant advancement in culturally-adaptive medical diagnosis through graph-theoretic modeling and multilingual natural language processing. The integration of cultural glossaries, bias adjustment mechanisms, and cluster entropy analysis provides robust diagnostic capabilities across diverse populations.

The framework's ability to detect rare diseases through entropy analysis represents a novel contribution to medical AI, with potential applications in global health initiatives and underserved populations.

Future work will explore federated learning approaches for privacy-preserving cultural adaptation and integration with electronic health records for longitudinal patient monitoring.

6. FUTURE RESEARCH DIRECTIONS AND CLINICAL VALIDATION PLAN

The Symptom Cascade Analyzer (SCA) offers a groundbreaking approach to culturally-adaptive medical diagnosis, but its current validation lacks real clinical data. To meet global health needs, particularly in underserved African populations, we outline future research and validation strategies.

First, we will expand the multilingual corpus to 25,000 symptom-disease pairs, incorporating dialects from 20 languages across sub-Saharan Africa, South Asia, and Latin America. This will refine the $\text{Gloss}(c,\ell)$ function, targeting a 30 percent improvement in accuracy for culturally-specific terms like "stomach fire."

Second, a pilot study with Mary Global Health will deploy SCA in 10 rural clinics in Kenya and Nigeria. Over 18 months, 1,000 patients will provide symptom descriptions, with diagnoses compared to local physician assessments. This will validate the 23 percent accuracy boost, using Cohen's kappa to ensure $\kappa > 0.85$ consistency across cultures.

Third, we will explore federated learning to preserve patient privacy [9], training SCA on decentralized data from multiple regions. This will enhance the GAT model's adaptability, aiming for a 15 percent increase in rare disease detection sensitivity.

Lastly, clinical validation will adhere to strict ethical standards. An IRB-approved protocol will secure informed consent, partnering with WHO to engage communities in study design. Metrics like diagnostic accuracy and patient satisfaction will be tracked over 24 months, targeting 95 percent reliability.

ACKNOWLEDGEMENTS

This work was assisted by Grok AI (xAI, 2025) for content generation and structuring.

CONFLICTS OF INTEREST

The author declares no conflicts of interest regarding the publication of this paper.

REFERENCES

1. Koleck, T.A., Dreisbach, C., Bourne, P.E. and Bakken, S. (2019) Natural Language Processing of Symptoms Documented in Free-Text Narratives of Electronic Health Records: A Systematic Review. *Journal of the American Medical Informatics Association*, **26**, 364-379. https://doi.org/10.1093/jamia/ocy173
2. Hossain, E., Rana, R., Higgins, N., Soar, J., Barua, P.D., Pisani, A.R., *et al.* (2023) Natural Language Processing in Electronic Health Records in Relation to Healthcare Decision-Making: A Systematic Review. *Computers in Biology and Medicine*, **155**, Article ID: 106649. https://doi.org/10.1016/j.compbiomed.2023.106649
3. Tahabi, F.M., Storey, S. and Luo, X. (2023) SymptomGraph: Identifying Symptom Clusters from Narrative Clinical Notes Using Graph Clustering. *Proceedings of the* 38*th ACM/SIGAPP Symposium on Applied Computing*, Tallinn Estonia, 27-31 March 2023, 518-527. https://doi.org/10.1145/3555776.3577685
4. Liu, C.C., Gurevych, I. and Korhonen, A. (2025) Culturally Aware and Adapted NLP: A Taxonomy and a Survey of the State of the Art. *Transactions of the Association for Computational Linguistics*, **13**, 652-689. https://doi.org/10.1162/tacl_a_00760
5. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., *et al.* (2020) Unsupervised Cross-Lingual Representation Learning at Scale. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, 5-10 July 2020, 8440-8451. https://doi.org/10.18653/v1/2020.acl-main.747
6. Jannat, A. (2025) Machine Learning Methods for Rare Disease Detection: A Systematic Review. Master's Thesis, University of Eastern Finland.
7. Swartz, J.B. (1998) An Entropy-Based Algorithm for Detecting Clusters of Cases and Controls and Its Comparison with a Method Using Nearest Neighbours. *Health & Place*, **4**, 67-77.

https://doi.org/10.1016/s1353-8292(97)00026-9

8. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P. and Bengio, Y. (2018) Graph Attention Networks. arXiv: 1710.10903.
9. McMahan, H.B., Moore, E., Ramage, D., Hampson, S. and Aguera y Arcas, B. (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. arXiv: 1602.05629.