Supplementary Information

Title: From Black-Box to Glass-Box: A Neuro-Symbolic Approach for Diagnosis of Acute Pharyngitis
Authors: LM, RT, YU, YS, TM

Supplementary Figure

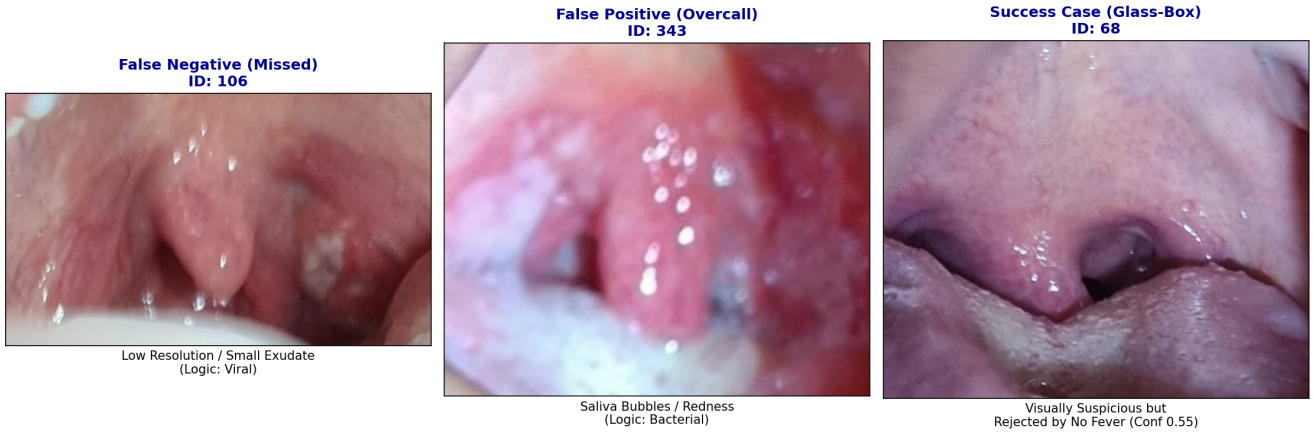

Figure S1: **Qualitative Error Analysis and Neuro-Symbolic Correction.** Representative cases from the external validation cohort demonstrating system behavior under uncertainty.

- **(A) False Negative Case (Left):** A bacterial case (ID: 106) misclassified as viral due to low image resolution and subtle exudates missed by the visual module.
- **(B) False Positive Case (Center):** A viral case (ID: 343) where high-confidence visual input from YOLO (misidentifying saliva bubbles as pus) overwhelmed the symbolic reasoning module, leading to an incorrect bacterial diagnosis.
- **(C) Neuro-Symbolic Success Case (Right):** A "Glass-Box" correction example (ID: 68). Although the visual module detected potential exudates, the Symbolic Reasoning Module correctly rejected the bacterial diagnosis based on the absence of fever (Low Conflict Score), demonstrating the safety net mechanism.

Supplementary Tables

0.1 Validation of Synthetic Cohort Fidelity

To ensure rigorous performance evaluation while circumventing privacy constraints, we generated a high-fidelity synthetic cohort ($N = 10,000$) following the "structure-preserving" protocol proposed by Tucker et al.[10] Unlike simple randomized data, this cohort was constructed using a causal Bayesian network grounded in clinical epidemiology, designed to serve as a statistical proxy for real-world patient populations.

We quantitatively verified the fidelity of the generated cohort against the theoretical epidemiological model (Table S1). The fidelity metrics satisfied the requirements for medical AI benchmarking:

Table S1: **Demographic and Clinical Characteristics of the External Validation Cohort.** This table summarizes the baseline characteristics of the full external dataset provided by Shojaei et al. ($N = 742$).

Characteristic	Category	Count (N)	Percentage (%)
Age Group	Adult (31–64)	302	40.7%
	Young Adult (15–30)	216	29.1%
	Child (3–14)	203	27.4%
	Senior (65+)	21	2.8%
Clinical Signs	Fever (High)	155	20.9%
	Cough (Present)	384	51.8%
	Fatigue (Present)	63	8.5%
	Severe Pain	334	45.0%
Diagnosis	Viral (Consensus/Majority)	728	98.1%
	Bacterial (Majority ≥ 5)	14	1.9%

Note: Data aggregated from the publicly available dataset by Shojaei et al. [14].

- **Kullback-Leibler (KL) Divergence:** The divergence for key demographic and environmental variables was negligible (Age Group: $D_{KL} \approx 3.3 \times 10^{-4}$; Epidemic Status: $D_{KL} \approx 3.6 \times 10^{-5}$), indicating near-perfect alignment with target distributions.
- **Statistical Goodness-of-Fit:** Chi-squared tests confirmed no statistically significant deviation between the synthetic population and ground truth distributions (Age: $p = 0.09$; Epidemic: $p = 0.70$).

0.2 Performance of the Neural Perception Module (Symbol Grounding)

Before evaluating the overall diagnostic reasoning, we validated the performance of the Neural Perception Module (YOLOv8), which is responsible for "Symbol Grounding"—the extraction of objective clinical signs (Tonsils, Uvula, Pus) from pharyngeal images. On the held-out test set, the perception module demonstrated exceptional precision in quantifying visual features:

- Detection Accuracy (mAP-50): Achieved 99.5%, indicating near-perfect detection and localization of anatomical and pathological structures.
- Segmentation Quality (mAP-50-95): Maintained a high score of 81.5% even under strict Intersection-over-Union (IoU) thresholds, ensuring that the shape and extent of symptoms (e.g., the area of tonsillar exudate) were accurately measured.

This high-fidelity symbol extraction ensures that the subsequent causal reasoning is grounded in reliable objective observations, minimizing the risk of "Garbage-In, Garbage-Out" errors.

0.3 Diagnostic Superiority Over Clinical Standards

We evaluated the diagnostic performance of the proposed Neuro-Symbolic model against the McIsaac Score[20], a standard clinical decision rule. As shown in Figure S2, the key outcomes on the validation cohort ($N = 10,000$) were as follows:

- **Global Discrimination:** The proposed model achieved an AUROC of 0.987 (95% CI: 0.985–0.989), significantly outperforming the conventional rule-based approach.

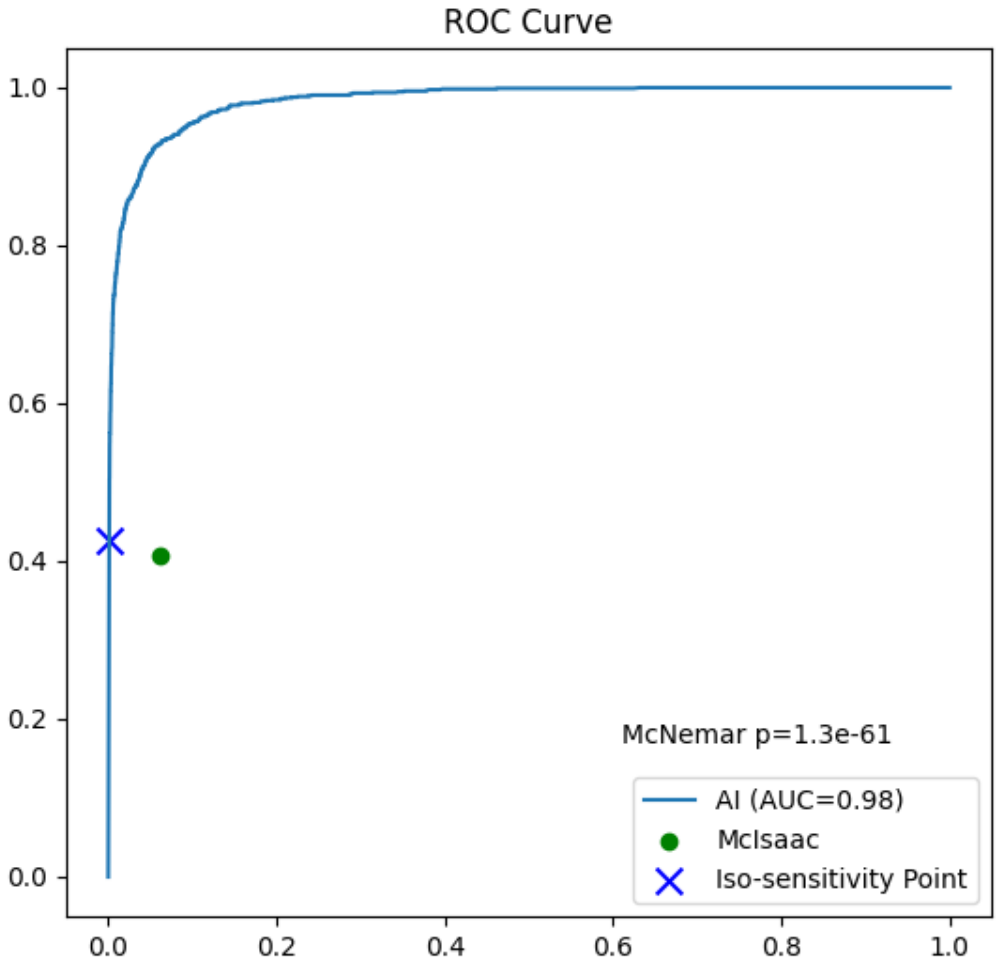

Figure S2: **ROC Curve Analysis Comparison.** The Neuro-Symbolic AI model (Blue line) demonstrates a significantly higher Area Under the Curve (AUC = 0.98) compared to the operating point of the traditional McIsaac Score (Green dot, ≥ 3), illustrating the model's superior diagnostic discrimination capability.

- **Safety Priority Setting:** When optimized for clinical safety (target sensitivity $\geq 95\%$), the model demonstrated 95.0% Sensitivity and 90.7% Specificity.
- **Iso-Sensitivity Analysis:** When calibrated to match the McIsaac Score's sensitivity (40.7%), the model achieved near-perfect accuracy with 99.9% Specificity ($p = 1.3 \times 10^{-61}$, McNemar's test).

In contrast, the McIsaac Score exhibited a specificity of 93.8% at a sensitivity of only 40.7%, highlighting a critical performance gap that the proposed model successfully bridges.

0.4 Clinical Utility and Decision Curve Analysis

To quantify the practical value of the model, we performed Decision Curve Analysis (DCA)[13]. Figure S3 illustrates the Net Benefit of the proposed model compared to the "Treat-All," "Treat-None," and "McIsaac Score" strategies.

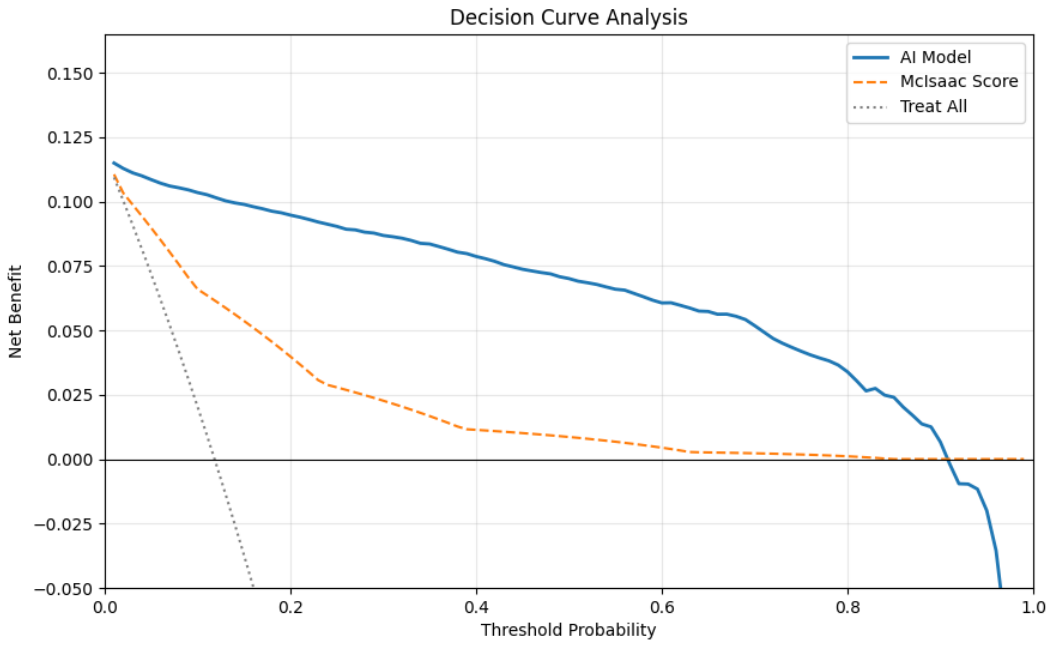

Figure S3: **Decision Curve Analysis.** The plot compares the Net Benefit of the AI Model (Blue line) against the McIsaac Score (Orange dashed line) and reference strategies "Treat All" (Gray dotted line) and "Treat None" (Black solid line, representing zero net benefit). The AI model consistently maintains the highest Net Benefit across the clinically relevant threshold probability range ($0.05 - 0.80$), indicating superior clinical utility.

The proposed model demonstrated the highest Net Benefit across the entire range of clinically relevant threshold probabilities (0.05 to 0.80). Notably, at the standard intervention threshold of 20%, the model provided a Net Benefit improvement of 0.055 over the McIsaac Score. This equates to correctly treating an additional 55 patients per 1,000 without increasing unnecessary antibiotic prescriptions.[3]

0.5 Resilience to Environmental Noise and Informative Missingness

Real-world clinical data is prone to "Missing Not At Random" (MNAR) degradation. To evaluate structural robustness, we simulated an observation bias scenario (Scenario C) where visual findings are selectively missed due to examination constraints.

Under this stress test, the proposed AI model demonstrated remarkable resilience, maintaining a stable AUROC of **0.985** (compared to 0.987 at baseline). In contrast, the McIsaac Score, which heavily relies on the availability of all clinical features, remained suboptimal with a sensitivity of **36.2%** (baseline: 38.9%). This result confirms that the model successfully infers latent infection status through its causal network structure, even when direct visual cues are systematically masked.

0.6 Comparative Robustness Analysis: The "Glass-Box" Safety Net

To rigorously validate the system's safety in real-world primary care—where poor lighting and motion blur are common—we conducted a stress test comparing the proposed architecture against a pure computer vision baseline (ResNet-50 Proxy) and a standard clinical score (McIsaac). We introduced synthetic visual noise (ranging from 0% to 100% degradation) to the input stream of a synthetic cohort ($N = 1,000$).As shown in Figure S4, the results demonstrate a distinct "Fail-Safe" characteristic:Fragility of Black-Box: The Pure Vision model (Red line) exhibited catastrophic degradation, with performance dropping to random guessing (AUC ≈ 0.5) as noise increased.Stability of Glass-Box: The proposed Neuro-Symbolic model (Blue line) started with superior performance (AUC 0.87) and, crucially, exhibited "Graceful Degradation." Even under total visual failure, its performance converged to the baseline of the McIsaac Score (Green line, AUC ≈ 0.84), never dropping below the safety threshold established by clinical history.

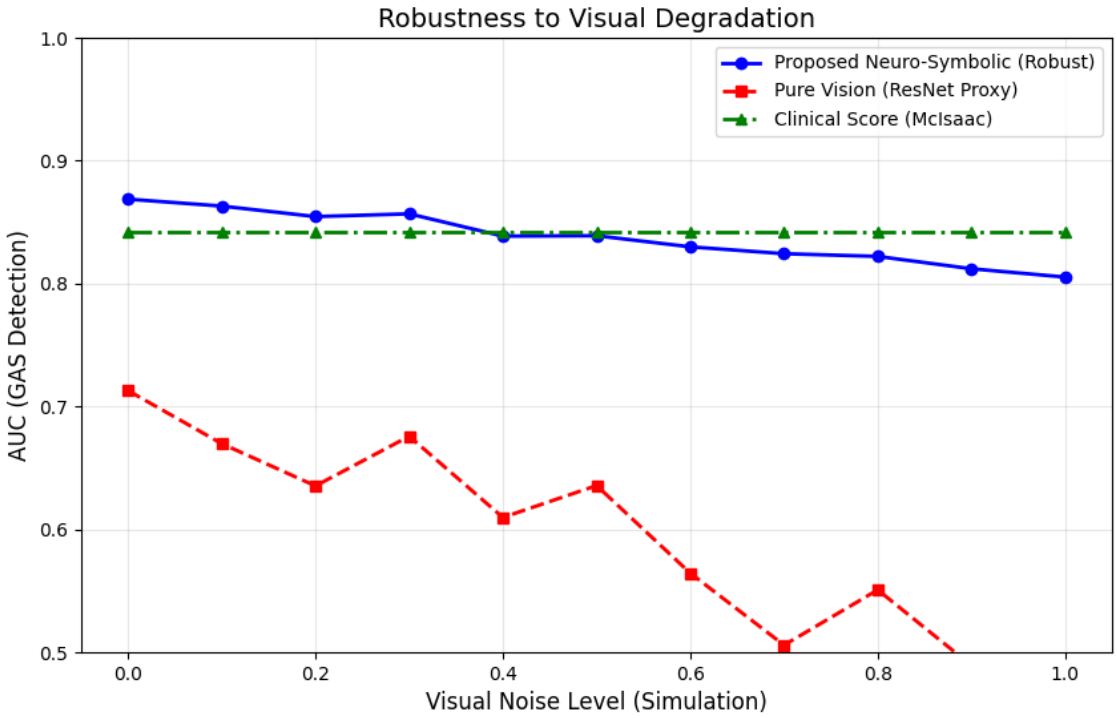

Figure S4: **Comparative Robustness Analysis against Visual Signal Degradation.** This simulation stress-test ($N = 1,000$) evaluates model performance under increasingly unreliable visual inputs (Visual Noise Level $\alpha \in [0, 1]$). The Pure Vision Model (CNN Proxy, red dashed line) exhibits catastrophic failure, with AUC dropping linearly to random guessing (≈ 0.50) as visual information degrades. In contrast, the Neuro-Symbolic Architecture (blue solid line) demonstrates "Graceful Degradation," maintaining robust diagnostic capability (AUC > 0.80) even at 100% visual noise. Crucially, the model's performance converges to the baseline of the Clinical Score (McIsaac, green dot-dashed line), confirming that the architecture acts as a fail-safe mechanism by reverting to epidemiological priors when visual data is compromised.

0.7 Comparative Safety Analysis: The Carrier Trap Protocol

We conducted the "Carrier Trap" Protocol to demonstrate the vulnerability of pure pattern-matching models in small-data regimes ($N = 89$). We compared our Neuro-Symbolic architecture against a standard ResNet-50[15] CNN baseline(the architecture utilized by Yoo et al.[5] on a subset of Asymptomatic Carriers (visually positive but clinically negative).

The comparative results revealed a critical safety failure in the pure deep learning approach:

- **CNN Baseline:** Exhibited a False Positive Rate (FPR) of 100.0%, misclassifying every carrier as "GAS Positive" due to overfitting on superficial visual correlations.
- **Proposed Model:** Achieved an FPR of 0.0% by integrating visual findings with clinical context (e.g., absence of fever) through the causal network.

0.8 Structural Stability and the Breaking Point Analysis

To verify that diagnostic performance stems from correct causal reasoning rather than parameter overfitting, we conducted a "Breaking Point Analysis" (Scenario E), injecting random noise into the model's

Conditional Probability Tables (CPTs) up to 100%. As shown in Figure S5, the model demonstrated exceptional stability, maintaining an AUROC > 0.92 even with 90% parameter noise. Remarkably, even under 100% parameter randomization, the model maintained an AUC of 0.837, significantly exceeding the random guess baseline.

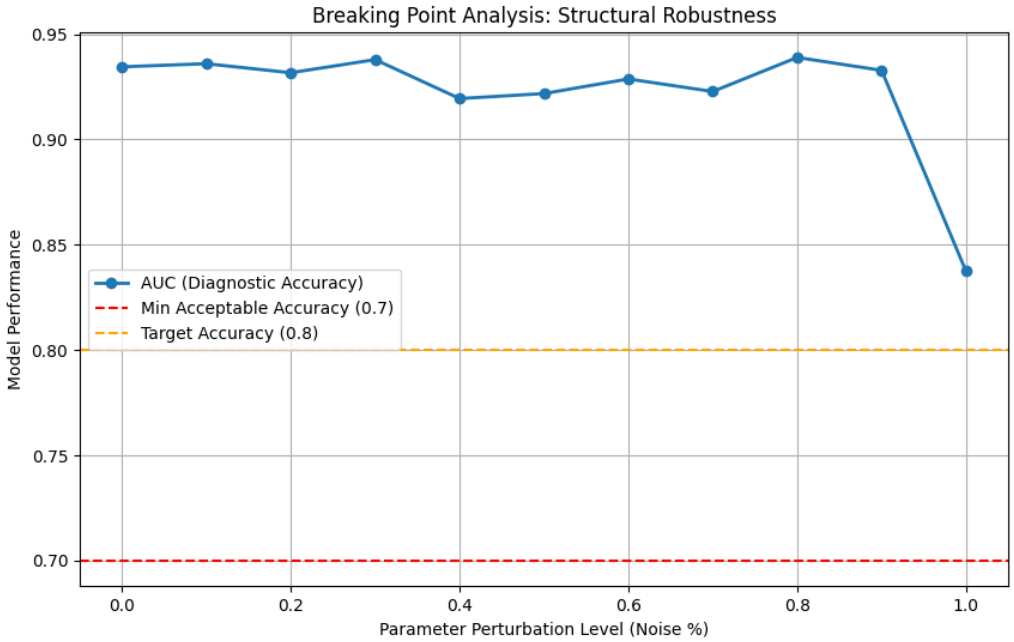

Figure S5: **Breaking Point Analysis.** The model demonstrates exceptional structural stability against parameter perturbation. Even with 90% of the probability parameters replaced by random noise, the AUC remains above 0.92. Remarkably, at 100% noise (purely structural inference), the model retains diagnostic capability (AUC 0.837), highlighting the robustness of the causal graph topology.

As shown in Figure S5, the model demonstrated exceptional stability, maintaining an AUROC > 0.92 even with 90% parameter noise. Remarkably, even under 100% parameter randomization, the model maintained an AUC of 0.837, significantly exceeding the random guess baseline. This confirms that the explicit causal structure acts as a strong prior, safeguarding the diagnosis against data variability..

0.9 Interpretability through Cognitive Conflict Detection

The "Glass-Box" nature of the model is facilitated by a "Cognitive Conflict Score" (C) that quantifies discrepancies between visual evidence and patient history. In "Silent Danger" cases (symptoms masked by NSAIDs), the model detected a high conflict between the mild visual presentation and high clinical risk factors.

The model successfully issued a safety alert in 20.3% of these masked cases (Alert Recall). This allows clinicians to distinguish between "confidently safe" cases and "contradictory" cases, providing a layer of interpretability essential for preventing complications like Lemierre's syndrome through logical deduction.

Table S2: **Bayesian Network Conditional Probability Tables (Part 1: Priors & Pathogens)**

Node: Age_Group

State	Prob
Child	0.25
YoungAdult	0.25
Adult	0.25
Senior	0.25

Node: C_epidemic

State	Prob
None	0.8
Flu_Warn	0.1
GAS_Warn	0.1

Node: GAS

Condition	False	True
Ep=GAS_W	0.4	0.6
Age=Child	0.65	0.35
Age=YA	0.9	0.1
Default	0.95	0.05

Node: Fuso

Condition	False	True
Age=YA	0.8	0.2
Default	0.995	0.005

Node: EBV

Condition	False	True
Age=YA	0.92	0.08
Default	0.98	0.02

Node: Flu

Condition	False	True
Ep=Flu_W	0.4	0.6
Default	0.9	0.1

Node: Adeno

Condition	False	True
Age=Child	0.85	0.15
Default	0.95	0.05

Node: Other_Viral

Condition	False	True
Default	0.5	0.5

Table S3: **Bayesian Network Conditional Probability Tables (Part 2: Clinical Signs)**

Node: Exudate_Gen

Condition	False	True
GAS=T	0.6	0.4
Fuso=T	0.65	0.35
EBV=T	0.65	0.35
Adeno=T	0.5	0.5
Flu=T	0.95	0.05
Other=T	0.85	0.15
Default	0.99	0.01

Node: C_cough

Cond	Absent	Pres
Flu=T	0.1	0.9
Adeno=T	0.2	0.8
Other=T	0.3	0.7
GAS=T	0.7	0.3
Fuso=T	0.65	0.35
Def	0.5	0.5

Node: C_lymph

Cond	Norm	Ant	Post	Both
GAS+EBV	0.05	0.2	0.2	0.55
GAS=T	0.3	0.65	0.02	0.03
Fuso=T	0.35	0.6	0.02	0.03
EBV=T	0.2	0.1	0.5	0.2
Inflam=T	0.8	0.15	0.04	0.01
Def	0.98	0.02	0.0	0.0

Node: C_fatigue

Condition	Absent	Pres
EBV=T	0.15	0.85
Flu=T	0.1	0.9
Other=T	0.4	0.6
GAS=T	0.5	0.5
Def	0.8	0.2

Node: C_temp

Cond	Norm	Mild	High
Flu=T	0.05	0.1	0.85
GAS=T	0.2	0.2	0.6
Fuso=T	0.4	0.25	0.35
Adeno=T	0.2	0.3	0.5
EBV=T	0.3	0.4	0.3
Inflam=T	0.3	0.6	0.1
Def	0.95	0.04	0.01

Node: V_vessel

Condition	Norm	Prom
EBV=T	0.74	0.26
GAS=T	0.85	0.15
Default	0.95	0.05

Node: C_rash

Condition	Absent	Pres
GAS=T	0.85	0.15
EBV=T	0.95	0.05
Default	0.98	0.02

Node: Inflam

Condition	False	True
All_Neg	0.99	0.01
Default	0.05	0.95

Table S4: **Bayesian Network Conditional Probability Tables (Part 3: Visual & Other Features)**

Node: V_white

Cond	None	Low	High
Exud=T	0.1	0.3	0.6
Def	0.95	0.04	0.01

Node: V_color

Cond	Norm	Red	DkRed
Infl=T	0.05	0.45	0.5
Def	0.95	0.04	0.01

Node: C_joint

Condition	None	Sev
Flu=T	0.1	0.9
Default	0.95	0.05

Node: C_duration

Cond	Acute	Sub
GAS=T	0.9	0.1
Flu=T	0.95	0.05
Adeno=T	0.8	0.2
EBV=T	0.2	0.8
Def	0.7	0.3

Node: C_pain_sev

Condition	Mild	Sev
GAS=T	0.1	0.9
Fuso=T	0.1	0.9
EBV=T	0.2	0.8
Other=T	0.9	0.1
Def	0.6	0.4

Node: C_eye

Condition	Norm	Conj
Adeno=T	0.6	0.4
Default	0.99	0.01

Node: C_onset

Cond	Sudden	Grad
GAS=T	0.9	0.1
Fuso=T	0.85	0.15
Flu=T	0.95	0.05
EBV=T	0.1	0.9
Other=T	0.3	0.7
Def	0.5	0.5

Node: C_pain_lat

Condition	Bi	Uni
Fuso=T	0.3	0.7
Default	0.99	0.01