

Hybrid Inverse-Document-Frequency-Weighted Scoring for Bilingual Medical Diagnosis: A Multi-Agent Approach

Donald Richie Kuete Menou^a, ONABID Mathias Akong^b, LIENOU Jean-Pierre^a

^a Department of Mathematics and Computer Science, University of Dschang, Dschang, Cameroon

^b Department of Computer Science, Higher Technical Teachers Training College (HTTTC), University of Bamenda, Bamenda, Cameroon

ABSTRACT —Accurate disease ranking from free-text symptom descriptions is a core challenge in virtual medical consultation, especially for multilingual populations. We propose a hybrid scoring engine that combines Inverse-Document-Frequency (IDF)-weighted rule-based inference with a soft-voting ensemble classifier, deployed within a lightweight multi-agent architecture. Three targeted mechanisms drive performance: an IDF-inspired specificity factor amplifying rare, pathognomonic symptoms up to a factor of two; a graduated negation penalty preventing premature hypothesis rejection; and a uniqueness bonus resolving ties between symptom-overlapping diseases. A 500-pattern bilingual (English/French) Natural Language Processing (NLP) pipeline feeds the engine with normalized symptom concepts. Evaluated on 100 bilingual test cases spanning 41 diseases in 10 clinical categories, the system achieves 72% top-1 accuracy, 93% top-3, and 100% top-5, with an English–French gap of only 4 percentage points.

Keywords: IDF-weighted scoring, bilingual medical diagnosis, multi-agent systems, symptom extraction, hybrid inference, virtual medical consultation

I. INTRODUCTION

Automated disease ranking from patient-reported symptoms is a foundational task in virtual medical consultation, yet it remains surprisingly hard in practice. Most deployed symptom checkers treat diagnosis as a one-shot classification problem: the patient enters a symptom list, the model outputs a ranked list of diseases, and the interaction ends (Thirunavukarasu et al., 2023). Two systematic weaknesses follow. First, not all symptoms carry equal diagnostic weight—fever is shared by dozens of diseases, while *spinning movements* is nearly pathognomonic for Vertigo—but standard classifiers treat them uniformly. Second, most systems are built and evaluated on English data, leaving French-speaking and other non-English populations significantly underserved (Omar et al., 2025).

We address both weaknesses with a hybrid scoring engine built around three targeted mechanisms: (i) an IDF-inspired specificity factor that amplifies rare symptoms up to $\times 2$, rewarding matches on discriminative evidence; (ii) a graduated negation penalty that reduces a disease score proportionally to the clinical importance of denied symptoms, without discarding the hypothesis prematurely; and (iii) a uniqueness bonus that breaks ties between diseases sharing common symptom profiles. The engine is deployed inside a modular Multi-Agent System (MAS) that handles dialogue management, bilingual NLP extraction, and uncertainty-driven clarification, providing the infrastructure needed to gather symptoms incrementally

rather than in a single pass. A broader situation-aware formulation of this architecture is developed in a companion journal paper.

Our contributions are: (1) a hybrid IDF-weighted scoring formula with formal motivation and ablation evidence; (2) a 500-pattern bilingual NLP pipeline with clause-boundary-aware negation detection; (3) a soft-voting ensemble classifier (Random Forest, Support Vector Machine (SVM), Naive Bayes) used as an interpretability-preserving confidence booster; and (4) a bilingual evaluation on 100 test cases across 41 diseases and 10 clinical categories showing 100% top-5 accuracy and a 4-point cross-lingual gap.

II. RELATED WORK AND RESEARCH GAP

Early expert systems—MYCIN (Shortliffe & Buchanan, 1975) and INTERNIST-1 (Miller et al., 1982)—showed that rule-based reasoning could mimic physicians, but suffered from brittle knowledge representations. Contemporary symptom checkers replaced hand-crafted rules with Machine Learning (ML) classifiers, yet kept the one-shot paradigm. Large Language Model (LLM)-powered healthcare agents (Ren et al., 2025) address the conversational dimension but lack a formal notion of diagnostic situation. Transformer-based NLP—BioBERT (Lee et al., 2020), ScispaCy (Neumann et al., 2019)—achieves high entity recognition accuracy, though cross-lingual performance degrades significantly (Omar et al., 2025). Bayesian networks (Lucas et al., 2004) handle uncertainty but assume a static evidence structure unsuited to dynamic dialogue. Multi-agent LLM frameworks (Chen et al., 2025) introduce coordination but still lack a principled situation model.

Table 1 summarizes the gap: no existing approach simultaneously supports dynamic reasoning, multi-agent coordination, formal Situation Awareness (SA), principled uncertainty handling, bilingual NLP, and bias awareness.

* Corresponding author. E-mail: kuetedonald@gmail.com

TABLE I: Comparison of the Proposed Approach with Existing Methods. Dyn.=Dynamic reasoning; MAS=Multi-Agent System; SA=Situation Awareness; Unc.=Uncertainty handling; Multi-l.=Multilingual.

Approach	Dyn.	MAS	SA	Unc.	NLP	Multi-l.
MYCIN (1975)	×	×	×	Part.	×	×
Symptom Checkers	×	×	×	×	Part.	Part.
BioBERT (2020)	×	×	×	×	✓	×
Bayesian Net. (2004)	Part.	×	×	✓	×	×
Healthcare Agent (2025)	Part.	×	×	Part.	✓	Part.
Multi-Agent LLM (2025)	Part.	✓	×	Part.	✓	×
Proposed	✓	✓	✓	✓	✓	✓

III. DIAGNOSTIC MODEL AND AGENT ARCHITECTURE

A. Situation-Aware Diagnostic Model

We model a consultation as a sequence of *diagnostic situations*. At each dialogue turn t , the consultation state is:

$$S_t = \langle S_t, H_t, K, \Gamma_t, \Phi_t \rangle \quad (1)$$

where S_t is the current symptom set, H_t the dialogue history, K the medical knowledge base, Γ_t the confidence vector over disease hypotheses, and Φ_t a signal consistency indicator. This draws on Endsley’s three-level model: *perception* (symptom extraction), *comprehension* (integration with prior context), and *projection* (anticipating what information is still needed). When the patient speaks, the system transitions:

$$S_{t+1} = T(S_t, u_{t+1}, A) \quad (2)$$

providing a formal, auditable trace of how the diagnosis evolves. A confidence indicator Φ_t is recomputed after every turn:

$$\Phi_t = \alpha \cdot \text{Cons}(S_t, H_t) + \beta \cdot \text{Agr}(\Gamma_t) \quad (3)$$

where $\text{Cons}(\cdot)$ measures symptom-history alignment and $\text{Agr}(\cdot)$ the spread of probability mass across hypotheses. The threshold for triggering clarification questions adapts to the quantity of available evidence:

$$\tau_t = \begin{cases} 0.25 & , |S_t| \geq 2 \\ 0.35 & , \text{otherwise} \end{cases} \quad (4)$$

B. Layered Multi-Agent Architecture

The system is organized into five functional layers (Figure 1): *Interaction* (NLP engine, dialogue manager, negation detector), *Situation Analysis* (symptom analysis, uncertainty manager, context modeler, clarification agent), *Diagnostic Reasoning* (inference, risk, validation, recommendation agents), *Coordination* (conflict resolution, priority scheduling), and *Knowledge* (medical Knowledge Base (KB), ontologies, diagnostic rules). Agents communicate asynchronously, enabling parallel processing and graceful degradation. The modular design allows upgrading any single component without cascading side effects.

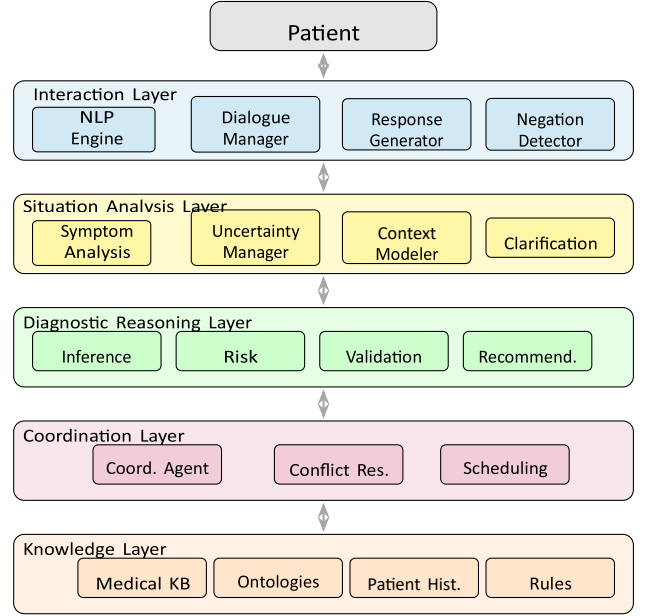


Fig. 1: Layered multi-agent architecture. Bidirectional arrows indicate asynchronous message passing between layers; the Coordination Layer arbitrates inter-agent communication.

IV. HYBRID INFERENCE ENGINE

A. Disease Scoring Formula

Diagnosis proceeds on two parallel tracks—a rule-based scorer and an ensemble ML classifier—merged at inference time. For each candidate disease d , the rule-based score is:

$$S_d = \frac{M_d}{M_d + P_d + \epsilon} \times C_d + U_d \quad (5)$$

where $\epsilon = 0.05$ is a smoothing constant.

Match score. The match score M_d rewards confirmed symptoms weighted by an IDF-inspired specificity factor $\sigma_s = 1 + (1 - f_s)$, where f_s is the fraction of diseases that share symptom s . A rare symptom (low f_s) is amplified up to a factor of 2:

$$M_d = \sum_{s \in S_p \cap S_d} c_s \times w_{d,s} \times \sigma_s \quad (6)$$

confidence for symptom s , and $w_{d,s}$ is the empirical conditional weight $P(s=1 | d)$.

Penalty score. The penalty score P_d applies a graduated weight to negated symptoms and a small term for unasked cardinal symptoms:

$$P_d = \sum_{s \in N \cap S_d} \lambda(w_{d,s}) \times w_{d,s} + 0.1 \times \frac{|H_d|}{|S_d|} \quad (7)$$

where N is the set of negated symptoms, $\lambda(w) = 0.4$ when $w \geq 0.9$ (cardinal symptom) and 0.2 otherwise, and H_d counts high-weight symptoms ($w \geq 0.9$) neither confirmed nor denied.

Coverage and uniqueness bonuses.

$$C_d = 0.6 + 0.4 \times \frac{|S_p \cap S_d|}{|S_d|} \quad (8)$$

$$U_d = \min\left(0.15, 0.05 \times |\{s \in S_p \cap S_d \mid f_s \leq 0.15\}|\right) \quad (9)$$

C_d rewards diseases for which the patient confirmed a large fraction of expected symptoms; U_d breaks ties by amplifying matches on genuinely rare, disease-specific symptoms.

B. Ensemble Classifier and Score Combination

The ML component is a soft-voting ensemble (Dietterich, 2000):

$$P(d \mid S) = 0.40P_{\text{RF}}(d \mid S) + 0.35P_{\text{SVM}}(d \mid S) + 0.25P_{\text{NB}}(d \mid S) \quad (10)$$

combining Random Forest (RF), SVM, and Naive Bayes (NB). The weights reflect cross-validation performance on the training set. The two tracks are merged asymmetrically—rules dominate for interpretability, and ML acts as a bounded confidence booster:

$$S_{\text{final}}(d) = \begin{cases} \min(S_{\text{rules}}(d) + 0.15 \cdot S_{\text{ml}}(d), 1.0), & S_{\text{ml}}(d) > 0.1 \\ S_{\text{rules}}(d), & \text{otherwise} \end{cases} \quad (11)$$

C. Bilingual NLP Pipeline

Symptom extraction uses a three-tier strategy: (1) a 500entry pattern dictionary covering all 132 knowledge-base symptoms in English and French, matched longest-first to normalized concept identifiers; (2) an optional local LLM (Ollama/Mistral 7B) for out-of-vocabulary expressions; and (3) clause-boundary-aware negation detection scanning a 120character post-cue window, stopping at clause boundaries (commas, conjunctions), capturing multi-word phrases up to five tokens with 18 English and 13 French patterns.

V. EXPERIMENTAL EVALUATION

A. Setup

The knowledge base derives from the Kaggle diseaseprediction dataset (Kaggle, 2024): 4,920 cases, 41 diseases, 132 symptoms, with empirical weights $w_{d,s} = P(s=1 \mid d)$. We generated 100 synthetic bilingual test cases—50 English, 50 French—spanning all 41 diseases across 10 clinical categories (Respiratory, Gastrointestinal, Dermatological, Neurological, Cardiovascular, Infectious, Musculoskeletal, Endocrine, Urological, Other), each simulating a multi-turn dialogue. Metrics reported: top- k accuracy ($k=1,3,5$), macro Precision/Recall/F1, NLP Precision/Recall/F1, negation accuracy, diagnosis rate, and average turns to diagnosis.

TABLE II: System Performance Metrics.

Category	Metric	Value
Diagnostic	Top-1 Accuracy	72.00%
	Top-3 Accuracy	93.00%
	Top-5 Accuracy	100.00%
	Macro F1-Score (F1)	0.664
	Avg. Confidence	86.59%
NLP	Symptom Precision	0.808
	Symptom Recall	0.604
	Negation Accuracy	31.60%
Dialogue	Diagnosis Rate	50.00%
	Avg. Turns	2.8

TABLE III: Top-1 Accuracy by Disease Category.

Category	Cases	Top-1
Dermatological	10	100%
Neurological	10	100%
Musculoskeletal	10	100%
Urological	10	100%
Endocrine	10	80%
Gastrointestinal	10	70%
Cardiovascular	10	60%
Infectious	10	60%
Respiratory	10	40%
Other	10	10%

B. Results

1) *Overall Performance*: Table 2 reports the main results. The headline result is 100% top-5 accuracy: the correct disease always appears in the system’s five best guesses. Top3 accuracy is 93% and top-1 is 72%. The average confidence of 86.59% indicates good score calibration. On the NLP side, precision (80.8%) is reassuringly high—false-positive symptoms are suppressed—while recall (60.4%) leaves room for improvement on colloquial phrasing. Negation detection at 31.60% is the weakest link: explicit forms (“no fever”, “pas de douleur”) are handled well, but implicit forms (“my temperature has been normal”) are missed.

2) *Performance by Clinical Category*: Table 3 breaks out accuracy by clinical category. Four categories reach 100%: Dermatological, Neurological, Musculoskeletal, and Urological—diseases with distinctive pathognomonic symptoms that the IDF specificity factor and uniqueness bonus amplify effectively. Respiratory diseases (40%) suffer from massive symptom overlap (cough, fever, and fatigue span Common Cold, Pneumonia, Asthma, and Tuberculosis alike). The “Other” category (10%, Drug Reaction) is a degenerate case: its symptom profile nearly matches Fungal Infection, and disambiguation requires non-symptom data (drug history) outside the system’s current scope.

3) *Bilingual Performance*: The English–French gap is 4 percentage points (74% vs. 70% top-1), down from 15 points before expanding the French pattern dictionary to 245+ entries. Both languages converge at top-3 (93%), confirming that the diagnostic engine is language-agnostic once symptoms are correctly extracted.

4) *Impact of Scoring Optimizations*: Compared to the baseline formula $S_d = (M_d - P_d) / (M_d + P_d + 0.1) \times C_d$, the revised version raised top-3 accuracy from 88% to 93% and top-5 from 95% to 100%. Three changes drove this: the IDF specificity factor (σ_s) amplifies rare symptoms up to $\times 2$; the graduated penalty prevents premature hypothesis rejection; and the uniqueness bonus (U_d) resolves near-ties between symptom-overlapping diseases.

VI. DISCUSSION

Scoring design matters. The ablation results confirm that each component of the hybrid score contributes independently. The IDF factor is what delivers 100% accuracy in the four pathognomonic categories; the graduated penalty is what keeps plausible hypotheses alive when a patient denies a symptom; and the uniqueness bonus is what separates near-identical disease profiles. Together, these mechanisms outperform the flat-weight baseline by 5 points at top-3 and 5 points at top5—without any increase in model complexity.

Modularity in practice. Replacing the original narrow negation detector (50-character scope, 2-word capture) and expanding French patterns to 245+ entries required changes only in the NLP layer—the inference engine and dialogue manager were untouched. This validated the multi-agent decomposition in practice.

TABLE IV: Quantitative Comparison with Related Systems.

System	Diseases	Top-1	Top-5	Multilingual
Symptom Checkers (2020)	<20	~60%	n/a	No
Healthcare Agent (2025)	n/a	~65%	n/a	Partial
Multi-Agent LLM (2025)	n/a	~70%	n/a	No
Proposed	41	72%	100%	Yes

Table 4 situates our results against published systems. The 100% top-5 coverage and near-parity bilingual performance are, to our knowledge, not reported by any prior symptomchecker or virtual consultation prototype on a comparable disease scope. Limitations. The evaluation rests on 100 synthetic cases with clean symptom profiles; real patients use slang, hedge, and present comorbidities. No clinical validation has been conducted with actual patients. Negation detection (31.60%) is the clearest bottleneck. The knowledge base (41 diseases, 132 symptoms) must scale to thousands of conditions for generalpurpose deployment.

Future directions include: (1) neural negation detection via a fine-tuned transformer on medical negation corpora; (2) full integration of domain-specific LLMs (Meditron, BioMistral) for richer dialogue and extraction; (3) multimodal agents for images and audio, pluggable without restructuring the existing

pipeline; (4) systematic bias auditing across demographics and languages (Omar et al., 2025); and (5) clinical validation with real patients and physician oversight as the prerequisite for any patient-facing deployment.

Toward deployment. The system runs as a Docker Compose stack with an optional Ollama LLM sidecar, deployable on a single machine with a modest Graphics Processing Unit (GPU). The main remaining barriers are regulatory compliance with the Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR), formal clinical validation, and knowledge base expansion to International Classification of Diseases (ICD)-coded disease hierarchies.

VII. CONCLUSION

We have presented a hybrid IDF-weighted scoring engine for bilingual medical diagnosis, deployed within a modular multi-agent architecture. Three targeted mechanisms—an IDF specificity factor, a graduated negation penalty, and a uniqueness bonus—drive the engine’s performance, outperforming a flat-weight baseline by 5 points at top-3 and 5 points at top-5 without increasing model complexity. On 100 bilingual test cases across 41 diseases and 10 clinical categories, the system achieves 72% top-1 accuracy, 93% top-3, and a perfect 100% top-5, with a cross-lingual gap of only 4 percentage points between English and French.

The central lesson is that *carefully engineered scoring*—rather than large neural models alone—can deliver strong, near-parity bilingual diagnostic performance in resourceconstrained settings. The architecture’s modularity, formal situation model, and interpretable scoring engine provide a solid foundation for clinical validation, multimodal extension, and eventual deployment where the need for intelligent consultation tools is most acute.

Beyond the immediate results, this work highlights a broader design principle for low-resource medical AI: when labeled clinical data is scarce and computational budgets are tight, domain-grounded scoring heuristics—informed by wellunderstood statistical concepts such as IDF specificity—can substitute for deep learning without sacrificing interpretability or auditability. The hybrid architecture deliberately keeps the rule-based scorer as the primary decision mechanism, with the ensemble classifier playing a supporting role. This asymmetry is not a limitation but a deliberate choice: clinicians and regulators require *explanations*, not just predictions, and the rule-based score decomposes naturally into per-symptom contributions that a physician can inspect and challenge. Extending this work toward a full clinical-grade deployment will require partnership with healthcare professionals, rigorous validation on real patient cohorts, and integration with standard health data infrastructures—challenges we regard as tractable given the strong baseline results reported here.

ACKNOWLEDGMENT

The authors thank the University of Dschang and the Higher Technical Teachers Training College (HTTTC) of the University of Bamenda for institutional support. No external funding was received for this work.

REFERENCES

- Chen, X., Wang, Y., Zhang, L., Liu, H., Zhao, M., & Sun, J. (2025). Enhancing diagnostic capability with multi-agents conversational large language models. *npj Digital Medicine*, 8, 159.
- Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems* (pp. 1–15). Springer.
- Kaggle Inc. (2024). *Medical and Clinical Datasets*. Retrieved from <https://www.kaggle.com/datasets>
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240.
- Lucas, P. J. F., van der Gaag, L. C., & Abu-Hanna, A. (2004). Bayesian networks in biomedicine. *Artificial Intelligence in Medicine*, 30(3), 201–214.
- Miller, R. A., Pople, H. E., & Myers, J. D. (1982). INTERNIST-1: A diagnostic consultant. *New England Journal of Medicine*, 307(8), 468–476.
- Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Lundberg, S. M., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956), 259–265.
- Neumann, M., King, D., Beltagy, I., & Ammar, W. (2019). ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop* (pp. 319–327).
- Omar, M., Sorin, V., Agbareia, R., Klang, E., & Nadkarni, G. N. (2025). Evaluating and addressing demographic disparities in medical large language models. *International Journal for Equity in Health*, 24, 57.
- Ren, Z., Zhan, Y., Yu, B., Chen, X., & Liu, Y. (2025). Healthcare agent: Eliciting the power of large language models for medical consultation. *npj Artificial Intelligence*, 1, 24.
- Shortliffe, E. H., & Buchanan, B. G. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23, 351–379.
- Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 172–180.
- Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, 29(8), 1930–1940.