Clinical Diagnosis of Rare Diseases using Leaky Noisy-OR Bayesian Networks

François ROUCOUX[a,1] and Sébastien JODOGNE[a]
[a] *Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), UCLouvain, 1348 Louvain-la-Neuve, Belgium*
ORCiD ID: François Roucoux https://orcid.org/0009-0007-1168-0688,
Sébastien Jodogne https://orcid.org/0000-0001-6685-7398

Abstract. This study presents a probabilistic method for the clinical diagnosis of rare diseases using leaky noisy-OR Bayesian networks automatically constructed from Orphanet and Human Phenotype Ontology data. The resulting model represents diseases and phenotypes as binary variables linked by causal probabilities derived from standardized annotations. Loopy belief propagation enables efficient approximate inference of disease posterior probabilities in large networks containing over 8,000 diseases and 9,000 finding variables. Evaluation on real clinical cases achieves 56.2% Top-3 diagnostic accuracy, in line with the reported performance of leading phenotype-based systems. The proposed framework demonstrates that interpretable and knowledge-grounded probabilistic reasoning can achieve state-of-the-art diagnostic performance for rare diseases while maintaining transparency and reproducibility. Unlike deep learning or ensemble models, it provides explicit causal explanations for each diagnostic hypothesis, enhancing interpretability and trustworthiness.

Keywords. Clinical diagnosis, Rare diseases, Leaky noisy-OR Bayesian networks, Orphanet, Human Phenotype Ontology

1. Introduction

In Europe, a disease is considered rare if it affects fewer than 1 in 2,000 individuals. Collectively, rare diseases affect 3.5–5.9% of the global population (263–446 million individuals) [1]. Their low prevalence limits clinical experience, often leading to prolonged diagnostic delays and a "diagnostic odyssey" involving multiple referrals and patient distress. Over 7,000 rare diseases have been described. Global initiatives such as Orphanet and OMIM compile clinical, epidemiological, and genetic data, and annotate diseases with Human Phenotype Ontology (HPO) terms to ensure interoperability [1,2,3].

Clinical decision support systems (CDSS) can help reduce diagnostic odyssey by assisting clinicians, hereby accelerating diagnosis [4]. Most rare disease CDSS rank diagnostic hypotheses by computing semantic or likelihood-based similarity between patient phenotypes (HPO terms) and disease profiles. Recent systems use ensemble

[1] Corresponding Author: François Roucoux, email: francois.roucoux@uclouvain.be
Research work carried out as part of the Clinom-X project, co-funded by the Walloon Region of Belgium and the European Union.

methods combining deep learning, semantic similarity, and probabilistic techniques to achieve state-of-the-art diagnostic accuracy [5].

This paper evaluates leaky noisy-OR networks, a subclass of Bayesian networks [6], as diagnostic engines for rare diseases. These models represent how multiple diseases independently contribute to phenotypes, providing an interpretable basis for probabilistic diagnosis [7]. The leaky extension accounts for unmodeled causes, improving robustness under incomplete knowledge. The specific structure of leaky noisy-OR networks enables computationally efficient inference on large domains. Results demonstrate that, populated with structured data from Orphanet and HPO, these models can be effectively applied to the diagnosis of rare diseases.

2. Methods

Our diagnostic engine consists of a two-layer Bayesian network comprising two types of binary variables: disease nodes $\mathbf{d} = \{d_1, \dots, d_n\}$ and finding (phenotype) nodes $\mathbf{f} = \{f_1, \dots, f_m\}$, as depicted in Figure 1. This well-established architecture captures the causal relationships between rare diseases and their observable manifestations. It is based on the following conditional independence assumptions: Diseases are *a priori* marginally independent, and findings are conditionally independent given the disease states. Such a network factorizes the joint probability distribution of diseases and findings as:

$$P(\mathbf{f}, \mathbf{d}) = P(\mathbf{f}|\mathbf{d})P(\mathbf{d}) = \left[\prod_{i=1}^{m} P(f_i|\mathbf{d})\right]\left[\prod_{j=1}^{n} P(d_j)\right].$$

The disease prior probabilities $P(d_j)$ used by our engine correspond to prevalence values derived from epidemiological data of Orphanet[2]. We further assume that the conditional probability $P(f_i|\mathbf{d})$ of each finding variable f_i follows a leaky noisy-OR model [7] (cf. Figure 2). Precisely, let $\boldsymbol{\pi}(f_i) \subseteq \{d_1, \dots, d_n\}$ denote the subset of diseases that can causally influence finding f_i. The probability that finding f_i is observed is defined as:

$$P(f_i = 1 \mid \mathbf{d}) = 1 - (1 - \lambda_i) \prod_{j \in \boldsymbol{\pi}(f_i)} (1 - p_{ij})^{d_j},$$

where $p_{ij} \in [0,1]$ denotes the causal activation probability of disease $d_j \in \{0,1\}$ on finding f_i, and $\lambda_i \in [0,1]$ represents the leak probability accounting for unmodeled causes. This specialized form reflects that multiple diseases can independently activate a finding, while the leak term accounts for unobserved etiologies. By assuming that diseases are independent and findings are conditionally independent, the model naturally handles multiple concurrent diseases and can diagnose conditions with overlapping phenotypes.

The actual value of the causal probability terms p_{ij} are obtained from the HPO ontology annotations, which provides the frequency of associations between rare diseases and the corresponding phenotypes[3]. The leak probability λ_i for each finding f_i is estimated heuristically as a logarithmic function of the number of diseases explicitly associated with f_i in the HPO annotations.

[2] https://www.orphadata.com/data/xml/en_product9_prev.xml (release of July 2025)

[3] https://github.com/obophenotype/human-phenotype-ontology/releases/tag/v2025-09-01

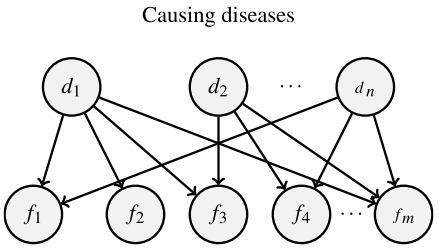

Figure 1. Two-layer Bayesian network with disease and finding binary variables. $d_j \in \{0,1\}$ indicates the absence or presence of disease j, and $f_i \in \{0,1\}$ indicates a negative or positive observation of finding i.

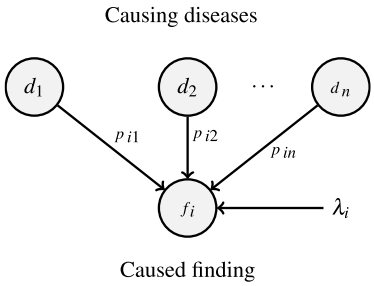

Figure 2. Leaky noisy-OR model where each finding f_i is activated independently by diseases $d_1, \ldots, d_n$ with causal probabilities p_{ij}. The leak parameter λ_i represents the probability of activation of the finding due to unmodeled causes.

Given the observed set of positive findings $\mathbf{f}^+ = \{f_i : f_i = 1\}$ and negative findings $\mathbf{f}^- = \{f_i : f_i = 0\}$, the diagnostic task consists in computing the posterior marginal probabilities $P(d_j | \mathbf{f}^+, \mathbf{f}^-)$ for each disease d_j. Exact posterior inference in leaky noisy-OR Bayesian networks is computationally infeasible for thousands of rare diseases and findings [8]. To achieve tractable inference while retaining diagnostic accuracy, we employ Loopy Belief Propagation (LBP), an approximate probabilistic inference algorithm [9]. LBP iteratively exchanges messages between connected nodes of the network, updating beliefs until convergence to an approximate posterior probability for each disease. Our implementation further employs asynchronous residual message scheduling to accelerate convergence [10] and applies damping to each message to mitigate oscillations during their update process [11].

3. Results

Our diagnostic engine was applied to a public dataset[4] of 384 real patients from published case reports with an established diagnosis of a rare disease [12]. Each case contains phenotypic HPO codes and at least one confirmed diagnostic OMIM code. The Bayesian network was generated in about 6 seconds using a non-optimized Python implementation. It contains 8,363 disease variables and 9,204 phenotype variables, as well as 156,168 directed causal links between them. Table 1 summarizes the accuracy of our model, with Top-K indicating that the K highest-ranked disease predictions inferred by the Bayesian network include a correct diagnostic OMIM code. Table 2 shows that high-confidence results (correct diagnosis with posterior probability $P \geq 0.9$) occurred in 45.9 % of cases. Table 3 compares the performance of our diagnostic engine with the accuracy reported for three state-of-the-art tools, demonstrating the competitiveness of our approach.

On average, the LBP algorithm converges after 35.9 iterations (range: 21−100). The mean inference time is 43.1 seconds per case, without specific optimization (range: 22.3−94.5 seconds). These timings were obtained on a standard CPU, showing that leaky noisy-OR Bayesian networks, when combined with an approximate inference algorithm, enable efficient handling of large-scale networks without specialized hardware.

[4] Benchmark dataset available at: https://zenodo.org/records/3905420 (version 1.0, Jun 24, 2020)

Table 1. Top-K diagnostic accuracy (n=384).

Top-K	Correct	Accuracy (%)
Top-1	179	46.6
Top-3	**216**	**56.2**
Top-5	232	60.4
Top-10	244	63.5
Top-100	318	82.8

Table 2. Posterior probabilities for correctly identified diagnoses present in Top-100 (n=318).

Posterior	Cases	Proportion (%)
$\mathbf{P \geq 0.9}$	**146**	**45.9**
$0.5 \leq P < 0.9$	23	7.2
$0.1 \leq P < 0.5$	47	14.8
$P < 0.1$	102	32.1

Table 3. Comparison of reported Top-3 diagnostic accuracy for phenotype-based diagnosis of rare diseases (accuracy values taken from reference publications).

Tool	Algorithm	Top-3 accuracy (%)
Our work	Leaky noisy-OR Bayesian network	**56.2**
PhenoBrain [5]	Ensemble of multiple methods (including deep learning)	51.3
LIRICAL [12]	Bayesian likelihood ratio	43.7
Phenomizer [13]	Ontology-based semantic similarity	35.3

4. Discussion

These experimental results demonstrate that our approach can achieve diagnostic performance comparable to leading phenotype-based systems for rare diseases. The Bayesian framework ensures that each link between a disease and a phenotype has a clear causal interpretation, enabling clinicians to trace how observations influence diagnostic probabilities. This transparency provides a decisive advantage over ensemble or deep learning-based methods (e.g., PhenoBrain): Unlike "black-box" models, our network supports the expert validation and regulatory auditability required for real-world clinical deployment. Furthermore, unlike systems such as LIRICAL or Phenomizer, our model naturally handles multi-morbidity by allowing multiple diseases to jointly explain a profile. This allows our engine to resolve complex cases with overlapping phenotypes where single-disease classifiers will fail.

Our diagnostic engine is released as a free and open-source software project[5] capable of processing patient cases encoded in the standard Phenopacket format to produce ranked OMIM diagnostic hypotheses with their posterior probabilities. By leveraging curated and widely recognized biomedical databases and ontologies, it provides a trustworthy foundation for diagnostic support. Future work will refine the estimation of leak probabilities, integrate temporal reasoning to capture symptom onset, and combine the engine with natural language processing to extract phenotypic information directly from electronic health records. Efforts will also focus on accelerating inference runtime, developing a user interface to improve explainability and usability, and establishing public benchmarks for fair comparison across diagnostic systems.

[5] Source code available at: https://forge.uclouvain.be/FrancoisRoucoux/rddnoisyornetwork

5. Conclusion

This paper investigates the application of leaky noisy-OR Bayesian networks to the diagnosis of rare diseases. We show that networks automatically derived from the Orphanet and HPO ontologies can effectively capture causal relationships between diseases and phenotypes. Bayesian networks offer the advantage of interpretability and support multi-morbidity inference, allowing multiple diseases to jointly explain observed phenotypes, which is an improvement over traditional single-disease approaches. Their probabilistic foundation ensures reproducible and explainable reasoning grounded in curated biomedical knowledge. The specific structure of leaky noisy-OR networks enables computational efficiency, allowing near real-time approximate inference using loopy belief propagation. An evaluation on 384 clinical cases achieved a Top-3 diagnostic accuracy of 56.2 %, which is comparable to state-of-the-art phenotype-based systems. Future work will refine leak estimation, incorporate temporal symptom data, build comparison benchmarks, and integrate the developed inference engine with electronic health records.

References

[1] Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet. 2020;28(2):165-73, doi:10.1038/s41431-019-0508-0.

[2] Hamosh A, Amberger JS, Bocchini C, Scott AF, Rasmussen SA. Online Mendelian Inheritance in Man (OMIM): Victor McKusick's magnum opus. Am J Med Genet A. 2021;185(11):3259-65, doi:10.1002/ajmg.a.62407.

[3] Gargano MA, Matentzoglu N, Coleman B, Addo-Lartey EB, Anagnostopoulos AV, Anderton J, et al. The Human Phenotype Ontology in 2024: phenotypes around the world. Nucleic Acids Res. 2024 Jan;52(D1):D1333-46, doi:10.1093/nar/gkad1005.

[4] Ronicke S, Hirsch MC, Türk E, Larionov K, Tientcheu D, Wagner AD. Can a decision support system accelerate rare disease diagnosis? Evaluating the potential impact of Ada DX in a retrospective study. Orphanet J Rare Dis. 2019 Mar;14(1):69, doi:10.1186/s13023-019-1040-6.

[5] Mao X, Huang Y, Jin Y, Wang L, Chen X, Liu H, et al. A phenotype-based AI pipeline outperforms human experts in differentially diagnosing rare diseases using EHRs. npj Digit Med. 2025;8:68, doi:10.1038/s41746-025-01452-1.

[6] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. San Mateo(CA): Morgan Kaufmann; 1988.

[7] Shwe MA, Middleton B, Heckerman DE, Henrion M, Horvitz EJ, Lehmann HP, et al. Probabilistic diagnosis using a reformulation of the INTERNIST-1/Quick Medical Reference (QMR) knowledge base. I. The probabilistic model and inference algorithms. Methods Inf Med. 1991;30(4):241-55.

[8] Jaakkola TS, Jordan MI. Variational probabilistic inference and the QMR-DT network. J Artif Intell Res. 1999;10:291-322, doi:10.1613/jair.594.

[9] Yedidia JS, Freeman WT, Weiss Y. Understanding belief propagation and its generalizations. Cambridge (MA): Massachusetts Institute of Technology, Artificial Intelligence Laboratory; 2001. Technical Report No.: AIM-2001-003.

[10] Elidan G, McGraw I, Koller D. Residual belief propagation: informed scheduling for message passing. In: Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI 2006); 2006. p. 165-73, doi:10.5555/3020419.3020440.

[11] Cohen L, Zivan R. Max-sum revisited: the real power of damping. In: Sukthankar G, Rodriguez-Aguilar JA, editors. Autonomous agents and multiagent systems. Cham (Switzerland): Springer International Publishing; 2017. p. 111-24, doi:10.1007/978-3-319-71679-4_8.

[12] Robinson PN, Ravanmehr V, Jacobsen JOB, Danis D, Zhang X, Carmody L, et al. Interpretable clinical genomics with a likelihood ratio paradigm. Am J Hum Genet. 2020;107(3):403-17, doi:10.1016/j.ajhg.2020.07.001.

[13] Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet. 2009;85(4):457-64, doi:10.1016/j.ajhg.2009.09.003.