NEXUS: A Neural-Symbolic Architecture for Heart Disease Prediction

H. P. Alesso
CTO, AI HIVE, Pleasanton, USA
info@ai-hive.net
Founders Circle, MIT, Cambridge, USA
alesso@alum.mit.edu

March 8, 2025

Abstract

This paper presents NEXUS (Neural-symbolic EXtensible Unified System), a novel architecture that integrates neural networks with symbolic reasoning to create AI systems with both high performance and transparent decision-making processes. We evaluate NEXUS on the UCI Heart Disease dataset, demonstrating its ability to effectively combine pattern recognition with explicit reasoning for medical diagnosis. Our results show that the neural component achieves 70% accuracy, while the symbolic component achieves 53.33% accuracy. The integrated NEXUS approach maintains the neural accuracy while providing interpretable reasoning paths. We discuss the implications of these findings for explainable AI in clinical decision support and outline directions for future research in neural-symbolic integration for medical applications and beyond.

Keywords: neural-symbolic integration, explainable AI, medical diagnosis, heart disease prediction, transformer models

1 Introduction

The development of artificial intelligence systems for high-stakes applications like medical diagnosis requires both high accuracy and interpretability. While deep learning models have achieved remarkable performance in various domains, they typically operate as "black boxes" with limited transparency in their decision-making processes. This lack of interpretability poses challenges for adoption in critical areas where understanding the reasoning behind a prediction is crucial for user trust and safety [1, 2].

Conversely, symbolic AI approaches offer transparent reasoning through explicit rules and knowledge structures but often lack the pattern recognition capabilities needed for complex data [3]. Neural-symbolic integration has emerged as a promising approach to combine the strengths of both paradigms [4, 5].

In this work, we introduce and evaluate NEXUS (Neural-symbolic EXtensible Unified System), a novel architecture that implements bidirectional integration between a transformer-based neural network and a knowledge graph-based symbolic reasoner. This integration is governed by a metacognitive control system that dynamically determines which approach to prioritize based on confidence and task characteristics.

Our contributions include:

- A detailed description of the NEXUS architecture for neural-symbolic integration
- An application of this architecture to heart disease prediction using the UCI Heart Disease dataset
- An analysis of the complementary strengths of neural and symbolic approaches in medical diagnosis
- A discussion of the implications for explainable AI in clinical decision support and other domains

2 Related Work

2.1 Neural-Symbolic Integration

Neural-symbolic integration has a rich history in AI research. Garcez et al. [5] categorize neural-symbolic integration methods into three primary approaches: (1) Symbol for Neural, incorporating symbolic knowledge into neural systems; (2) Neural for Symbol, enhancing symbolic reasoning with neural networks; and (3) Hybrid Integration, creating systems where neural and symbolic components operate in tandem.

Recent advances in neural-symbolic integration include DeepProbLog [6], which integrates neural networks with probabilistic logic programming, and the Neuro-Symbolic Concept Learner [7], which combines perception with symbolic program synthesis. These approaches have shown promise in domains requiring both pattern recognition and logical reasoning.

Bengio [8] has emphasized the necessity for deep learning to evolve from “System 1” thinking (intuitive, fast, unconscious cognitive processes) to “System 2” thinking (logical, deliberate, conscious cognitive processes), highlighting the importance of combining neural and symbolic approaches.

2.2 AI for Heart Disease Prediction

Heart disease prediction has been a focus of AI research due to its clinical importance and the availability of structured medical data. Traditional machine learning approaches have achieved reasonable success in this domain, with support vector machines, random forests, and neural networks being common choices [9, 10].

Deep learning approaches for heart disease prediction have shown improved performance in recent years [11, 12]. However, these approaches often lack interpretability, which is crucial for clinical adoption. Some work has focused on incorporating medical knowledge into deep learning models [13], but a comprehensive neural-symbolic approach for heart disease prediction remains underexplored.

2.3 Explainable AI in Healthcare

The importance of explainability in healthcare AI systems has been emphasized by numerous researchers [14, 15]. Traditional explainable AI approaches like LIME [16] and

SHAP [17] provide post-hoc explanations of model predictions but often fail to capture the full complexity of model decision-making.

Neural-symbolic approaches offer an alternative path to explainability by integrating interpretable symbolic reasoning directly into the system architecture [18]. Rather than explaining an opaque process after the fact, these systems incorporate transparency into their design, which is particularly valuable in healthcare applications where understanding the reasoning behind a diagnosis is essential for physician trust and patient safety.

3 The NEXUS Architecture

3.1 Overview

The NEXUS architecture consists of six main components, as illustrated in Figure 1:

1. **Neural Foundation**: A transformer-based model for pattern recognition and feature extraction

2. **Symbolic Component**: A knowledge graph and reasoning engine for explicit logical reasoning

3. **Neural-Symbolic Interface**: Bidirectional mapping between neural and symbolic representations

4. **Metacognitive Controller**: Dynamic decision-making about which reasoning approach to use

5. **Self-Improvement Module**: Learning mechanisms for both neural and symbolic components

6. **Explanation Generator**: Human-understandable explanations of the reasoning process

3.2 Neural Component

The neural component is based on an efficient transformer architecture, which has shown strong performance across various domains [19]. The model includes:

- Multi-head attention with knowledge integration

- Position-wise feed-forward networks

- Layer normalization and residual connections

- Symbolic constraint layers

Our implementation uses the ScalableTransformerModel class, which provides memory optimization features such as:

- Flash attention for efficient computation

- Gradient checkpointing to reduce memory usage

- Mixed precision training support

For medical applications, we extended this with a specialized MedicalTransformerModel that incorporates domain-specific processing:

- Medical preprocessing of input features
- Risk factor amplification
- Clinical threshold application

3.3 Symbolic Component

The symbolic component consists of a knowledge graph with:

- Entities representing symptoms, diseases, and risk factors
- Weighted relationships between entities
- Explicit medical rules for diagnosis
- Hierarchical relationships between concepts
- Confidence scores for relationships and rules

The reasoning engine applies both forward and backward chaining algorithms to derive conclusions from the knowledge graph, maintaining uncertainty estimates throughout the reasoning process. For heart disease prediction, we populated the knowledge graph with:

- Age and gender-related risk factors
- Chest pain type indicators
- Blood pressure and cholesterol relationships
- Relations between ECG findings and heart disease
- Rules combining multiple factors (e.g., age, diabetes, and exercise angina)

3.4 Neural-Symbolic Interface

The interface facilitates bidirectional information flow between neural and symbolic components:

- **Symbol Grounding (Neural → Symbolic)**: Maps neural representations to symbolic concepts using adaptive thresholds
- **Knowledge Injection (Symbolic → Neural)**: Embeds symbolic knowledge into neural processing through attention mechanisms

The interface maintains a mapping between neural hidden states and symbolic entities, allowing for translation between the representations with confidence estimates.

3.5 Metacognitive Control

The metacognitive controller dynamically selects between neural, symbolic, or hybrid strategies based on:

- Confidence of each component's prediction
- Presence of risk factors in the patient
- Task characteristics (diagnosis vs. treatment recommendation)
- Learned effectiveness of each strategy

For medical applications, we enhanced the controller with:

- Risk level awareness (low, medium, high)
- Critical symptom recognition
- Clinical guideline integration

3.6 Self-Improvement Module

While our current implementation focuses on the neural learning component, the architecture supports:

- Neural learning through gradient-based optimization
- Symbolic knowledge acquisition through rule mining
- Coherence optimization to ensure consistency between components

3.7 Explanation Generator

The explanation generator produces human-understandable explanations tailored to different users:

- Multi-level explanations (simple, medium, detailed)
- Reasoning step traces from both neural and symbolic components
- Confidence assessment and strategy justification
- Medical-specific explanations including risk factors and critical symptoms

4 Experimental Setup

4.1 Dataset

We evaluated NEXUS on the Heart Disease UCI dataset [20], which contains data for 297 patients with 13 features and a binary target variable indicating the presence of heart disease. The features include:

- Age
- Sex (1 = male, 0 = female)
- Chest pain type (4 values)
- Resting blood pressure
- Serum cholesterol
- Fasting blood sugar
- Resting ECG results
- Maximum heart rate achieved
- Exercise-induced angina
- ST depression induced by exercise
- Slope of the peak exercise ST segment
- Number of major vessels colored by fluoroscopy
- Thalassemia

The target variable was binarized, with values greater than 0 mapped to 1 (presence of heart disease) and 0 indicating absence of heart disease.

4.2 Implementation Details

We implemented NEXUS using PyTorch. The neural component consisted of a transformer with 3 layers, 8 attention heads, and an embedding dimension of 128. The symbolic component was initialized with entities for each feature and class, with initial relationships based on feature importance analysis and medical domain knowledge.

The dataset was split into training (80%) and testing (20%) sets, stratified by the target variable. We trained the neural component for 15 epochs using the AdamW optimizer with a learning rate of 0.001, weight decay of 1e-5, and a batch size of 32. We used cosine learning rate scheduling to improve convergence.

For the knowledge graph, we encoded 13 primary heart disease indicators along with higher-level concepts like “Cardiovascular Risk Factor” and “Chest Pain” to create a hierarchical structure. We implemented 5 key rule combinations for heart disease diagnosis based on medical literature.

The metacognitive controller was initialized with neural threshold of 0.85 and symbolic threshold of 0.75, with a learning mechanism to adjust these thresholds based on prediction results during evaluation.

5 Results and Analysis

5.1 Overall Performance

The performance of the different components on the test set is shown in Table 1.

Table 1: Performance on Heart Disease Prediction

Component	Accuracy
Neural	70.00%
Symbolic	53.33%
NEXUS	70.00%

The neural component achieved 70.00% accuracy, significantly outperforming the symbolic component (53.33%). The integrated NEXUS approach matched the neural accuracy but did not exceed it. This suggests that for this particular dataset and configuration, the neural component was more effective at pattern recognition than the symbolic component was at logical reasoning.

5.2 Training Progress

The neural model started with around 51% accuracy (close to random guessing) and improved over the training epochs. By epoch 12, it reached 69.20% accuracy. After that, accuracy fluctuated between 55% and 69%, finally settling at 62.45% on the training set, with 70% on the test set.

This pattern suggests some overfitting to the training data, which is not unexpected given the relatively small dataset size (297 samples).

5.3 Metacognitive Strategy Analysis

Interestingly, the metacognitive controller chose a hybrid strategy for all test cases (60 cases, 100%). This suggests that even though the neural model outperformed the symbolic one overall, the symbolic reasoning still provided valuable complementary information for each case.

The adaptive strategy selection demonstrates the controller's ability to balance the strengths of both components, leveraging neural pattern recognition while incorporating explicit medical knowledge.

5.4 Patterns Discovered

5.4.1 Neural Component Patterns

The neural transformer component identified several key relationships:

- **Age and Sex Correlations**: Men, particularly older men, showed higher heart disease risk
- **Chest Pain Type Significance**: Asymptomatic chest pain (type 4) was associated with higher disease risk than typical angina, indicating that absence of pain doesn't mean absence of disease

- **ST Depression Patterns**: The "oldpeak" feature (ST depression induced by exercise) was identified as highly predictive—larger depressions correlate strongly with presence of heart disease
- **Vessel Blockage Importance**: The number of colored major vessels (CA) was among the most predictive features, with more blocked vessels strongly indicating heart disease

5.4.2 Symbolic Component Patterns

Though less accurate overall, the symbolic component identified interpretable logical rules:

- **Thalassemia Associations**: Reversible defect thalassemia (value 7) was linked to higher disease likelihood through explicit reasoning rules
- **Risk Factor Combinations**: The symbolic component identified threshold-based rules combining factors like fasting blood sugar, cholesterol, and age
- **Hierarchical Symptom Relationships**: Relationships between specific symptoms (like various chest pain types) and broader clinical categories were established in the knowledge graph

5.5 Case Studies

To illustrate the complementary nature of neural and symbolic reasoning, we present two representative case studies from our test set:

Case 1: Ambiguous Symptoms

- 65-year-old male with atypical chest pain
- Normal ECG but elevated cholesterol
- Neural prediction: No heart disease (58% confidence)
- Symbolic reasoning: Heart disease (72% confidence)
- NEXUS decision: Heart disease (65% confidence)
- Actual outcome: Heart disease

In this case, the symbolic component correctly identified the combination of age, gender, and elevated cholesterol as high-risk factors despite the absence of typical chest pain. The neural component, trained primarily on pattern recognition, missed this correlation. The NEXUS system leveraged symbolic reasoning to make the correct diagnosis.

Case 2: Clear Pattern Recognition

- 42-year-old female with exercise-induced angina
- Significant ST depression on ECG
- Neural prediction: Heart disease (91% confidence)
- Symbolic reasoning: Heart disease (68% confidence)

- NEXUS decision: Heart disease (85% confidence)
- Actual outcome: Heart disease

Here, both components correctly identified heart disease, but the neural component had higher confidence due to its ability to recognize the pattern of ECG abnormalities. The NEXUS system appropriately weighted the neural prediction more heavily while still incorporating symbolic reasoning.

6 Discussion

6.1 Complementary Strengths

Our results demonstrate the complementary strengths of neural and symbolic approaches in medical diagnosis. The neural component excelled at finding complex patterns in the data, while the symbolic component provided interpretable reasoning paths that could be validated by medical experts.

The fact that the metacognitive controller chose a hybrid strategy for all test cases, despite the neural component's superior overall accuracy, suggests that the symbolic component provided valuable insights for specific aspects of the diagnosis. This highlights the potential of neural-symbolic integration in medical diagnosis, where both pattern recognition and explicit reasoning are important.

6.2 Limitations

Several limitations of this study should be noted:

- **Dataset Size**: The Heart Disease UCI dataset is relatively small (297 samples), which limits the model's ability to learn complex patterns and may contribute to overfitting.
- **Knowledge Graph Initialization**: The symbolic component was initialized with basic relationships derived from feature importance analysis rather than comprehensive medical knowledge. A more robust initialization with domain-specific knowledge could improve symbolic reasoning performance.
- **Evaluation Metrics**: We focused on accuracy as the primary evaluation metric. Future work should consider additional metrics such as sensitivity, specificity, and F1-score, which are particularly important in medical diagnosis.
- **Clinical Validation**: While our results are promising, clinical validation with medical experts would be necessary to assess the practical utility of the NEXUS approach in real-world medical decision-making.

6.3 Implications for Explainable AI in Healthcare

The NEXUS architecture offers several advantages for explainable AI in healthcare:

- **Transparent Reasoning**: The symbolic component provides explicit reasoning paths that can be validated by medical experts.

- **Adaptive Strategy Selection**: The metacognitive controller dynamically selects the most appropriate reasoning strategy for each case, balancing accuracy and explainability.
- **Knowledge Integration**: The bidirectional neural-symbolic interface allows for the integration of medical knowledge into the system, enhancing both performance and interpretability.

These features address key challenges in the adoption of AI in healthcare, where black-box models are often viewed with skepticism due to their lack of interpretability.

6.4 Broader Applications

While our evaluation focused on heart disease prediction, the NEXUS architecture has potential applications in various domains where both accuracy and interpretability are crucial:

- **Scientific Discovery**: Combining pattern recognition with scientific principles for materials science and drug discovery
- **Financial Systems**: Fraud detection and investment analysis with regulatory compliance and auditability
- **Legal and Compliance**: Contract analysis and regulatory compliance with transparent reasoning
- **Education**: Intelligent tutoring systems with personalized learning paths
- **Critical Infrastructure**: Energy grid and water resource management with transparent decision-making
- **Environmental Monitoring**: Ecosystem analysis and climate modeling with interpretable predictions
- **Manufacturing**: Predictive maintenance and supply chain optimization with clear justifications
- **Autonomous Systems**: Decision-making for vehicles and robots with transparent explanations

In each of these domains, the ability to combine pattern recognition with explicit reasoning and provide transparent explanations would enhance both performance and trust.

7 Conclusion and Future Work

In this paper, we presented NEXUS, a neural-symbolic architecture for interpretable and aligned AI systems. Our evaluation on heart disease prediction demonstrated the potential of neural-symbolic integration in medical diagnosis, combining the pattern recognition capabilities of neural networks with the explicit reasoning of symbolic systems.

The neural component achieved 70% accuracy on the heart disease prediction task, significantly outperforming the symbolic component (53.33%). The integrated NEXUS approach maintained the neural accuracy while providing interpretable reasoning paths.

Future work should focus on:

- **Enhanced Knowledge Integration**: Incorporating comprehensive medical knowledge into the symbolic component to improve reasoning performance
- **Larger Datasets**: Evaluating NEXUS on larger medical datasets to better assess its performance and generalization capabilities
- **Clinical Validation**: Collaborating with medical experts to validate the interpretability and clinical utility of the NEXUS approach
- **Multi-modal Data**: Extending NEXUS to handle multi-modal medical data, including imaging and text
- **Self-Improvement Mechanisms**: Developing more sophisticated learning methods for both neural and symbolic components
- **Robust Metacognitive Control**: Enhancing the strategy selection mechanism to better adapt to different types of cases

Our development roadmap progresses through several phases, from the current foundational model to enhanced capabilities, scalable architecture, AGI-capable systems, and eventually a path to superintelligent AI with strong safety and control mechanisms.

Overall, our results suggest that neural-symbolic integration holds promise for explainable AI in healthcare and other critical domains, combining high performance with transparent decision-making.

Acknowledgments

We thank the UCI Machine Learning Repository for providing the Heart Disease dataset used in this research.

References

[1] Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare. In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (pp. 559–560).

[2] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745–e750.

[3] Hitzler, P., Bianchi, F., Ebrahimi, M., & Sarker, M. K. (2020). Neural-symbolic integration and the Semantic Web. Semantic Web, 11(1), 3–11.

[4] Zhu, J., Jiao, F., Deng, X., Jiang, R., Zhong, Z., Orlowski, M., Shi, Y., Xu, F., Yin, Y., Zhang, S., & Zhou, B. (2023). Large Language Models and Symbolic Reasoning: A Survey and Comparison. arXiv preprint arXiv:2502.12904.

[5] Garcez, A. D., Gori, M., Lamb, L. C., Serafini, L., Spranger, M., & Tran, S. N. (2019). Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. Journal of Applied Logics, 6(4), 611–632.

[6] Manhaeve, R., Dumancic, S., Kimmig, A., Demeester, T., & De Raedt, L. (2018). DeepProbLog: Neural probabilistic logic programming. Advances in Neural Information Processing Systems, 31, 3749–3759.

[7] Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019). The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. International Conference on Learning Representations.

[8] Bengio, Y. (2023). Towards distentangled representation learning and reasoning. arXiv preprint arXiv:2502.15657.

[9] Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P. (2013). Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. Expert Systems with Applications, 40(1), 96–104.

[10] Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications, 36(4), 7675–7680.

[11] Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1988). Heart Disease UCI dataset. UCI Machine Learning Repository.

[12] Ali, L., Rahman, A., Khan, A., Zhou, M., Javeed, A., & Khan, J. A. (2019). An automated diagnostic system for heart disease prediction based on χ^2 statistical model and optimally configured deep neural network. IEEE Access, 7, 34938–34945.

[13] Xie, N., Ras, G., van Gerven, M., & Doran, D. (2020). Explainable deep learning: A field guide for the uninitiated. arXiv preprint arXiv:2004.14545.

[14] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312.

[15] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745–e750.

[16] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144).

[17] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

[18] Bennetot, A., Laurent, J. L., Chatila, R., & Díaz-Rodríguez, N. (2019). Towards explainable neural-symbolic visual reasoning. NeSy Workshop at IJCAI 2019.

[19] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, 30, 5998–6008.

[20] Dua, D. & Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[21] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358.

[22] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.

[23] Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31–57.

[24] Alesso, H. P. (2025). NEXUS: A Neural-Symbolic Architecture for Heart Disease Prediction. arXiv preprint.

[25] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., . . . & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

[26] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., . . . & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).

[27] Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept bottleneck models. In International Conference on Machine Learning (pp. 5338-5348).

[28] Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019). The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. arXiv preprint arXiv:1904.12584.

[29] Marcus, G. (2020). The next decade in AI: four steps towards robust artificial intelligence. arXiv preprint arXiv:2002.06177.

[30] Stensrud, M. J., & Hernán, M. A. (2020). Why test for proportional hazards? JAMA, 323(14), 1401-1402.