Medical Diagnosis and Risk Scoring using Bayesian Networks

Muhammad Asad [1]

Abstract

This study provides a comprehensive implementation and evaluation of Bayesian Networks for medical diagnosis, with specific applications to heart disease and diabetes classification. Unlike traditional discriminative models that learn direct mappings from features to outcomes, Bayesian Networks offer a probabilistic framework that models the joint distribution between attributes. We implement and optimize Bayesian Networks for two critical healthcare challenges: the UCI Heart Disease dataset and the Pima Indians Diabetes dataset, comparing performance against established baseline methods-Logistic Regression and XGBoost.

Our Bayesian Network implementation addresses key challenges including data discretization strategies, structure and parameter learning, inference, prediction, and evaluation. The diabetes dataset had binary classification but for the Heart Disease dataset, we explore both multiclass and binary classification. Our results demonstrate that Bayesian Networks, while computationally more expensive than discriminative models, provide a critical advantage on explainability and understanding as compared to other machine learning models and Neural networks while performing similar to these models.

1. Introduction

1.1. Motivation and Background

Cardiovascular diseases and diabetes mellitus represent two of the most pressing global health challenges, collectively affecting hundreds of millions of people worldwide. According to WHO, over 17.9 million deaths globally are due to cardiovascular diseases. Moreover, heart disease is the 2nd leading cause of death in Canada according to Canada.ca. Furthermore, diabetes prevalence has nearly quadrupled since 1980, reaching over 422 million adults globally. Early, accurate diagnosis of these conditions is paramount for effective intervention, disease management, and improved patient outcomes.

Traditional diagnostic approaches in medicine rely on clinical expertise, established guidelines, and manual interpretation of test results. While effective, these methods face inherent limitations: they are time-consuming, potentially subject to human cognitive biases, and may struggle to capture complex, multivariate interactions among numerous risk factors. The rise of electronic health records and exponential growth in medical data availability have created unprecedented opportunities to augment clinical decision-making with sophisticated data-driven approaches.

However, the application of machine learning in healthcare extends beyond mere predictive accuracy. Medical practitioners and regulatory bodies demand models that can explain their reasoning as well. Black-box models, despite high accuracy, face significant adoption barriers in clinical practice due to interpretability concerns, liability issues, and the fundamental need for physicians to understand and validate diagnostic recommendations.

Bayesian Networks emerge as a particularly promising approach for medical diagnosis, addressing many of these critical requirements. As probabilistic graphical models, Bayesian Networks represent variables as nodes and probabilistic dependencies as directed edges, explicitly modelling the causal structure underlying disease processes. Unlike discriminative models (such as Logistic Regression or XGBoost), that learn only the conditional probability P(Disease—Symptoms), Bayesian Networks model the full joint probability distribution P(Disease, Symptoms), enabling **Probabilistic Reasoning**, **Causal Interpretation Domain Knowledge Integration**, and **Explainability**.

1.2. Problem Statement and Research Focus

This research focuses on the implementation, optimization, and evaluation of Bayesian Networks for binary and multi medical classification tasks, specifically addressing two critical diagnostic challenges i.e. heart disease detection and diabetes diagnosis.

[1]Department of Electrical Engineering & Computer Science, York University, Ontario, Canada. Correspondence to: Muhammad Asad <mohdasad@yorku.ca>.

Project Report for EECS6327 (Fall 2025), Department of Electrical Engineering and Computer Science, York University.

This research investigates whether Bayesian Networks, despite their computational complexity and discretization requirements, achieve competitive predictive performance compared to already established discriminative models. Therefore, to establish the rigorous performance benchmarks, we've implemented two baselines methods i.e. Logistic regression and XGBoost models.

1.3. Report structure and sections

Starting off with the introduction, where we have mentioned our motivation, background, problem statement, and research objectives; the subsequent sections will be as follows: section 2 will focus on literature review followed by section 3, which will focus on methodology. Results will be displayed in section 4 followed by the discussion in section 5 and conclusion in section 6. References are presented at the end of this paper.

2. Literature Review

2.1. Machine Learning in Medical Diagnosis

Machine learning has been extensively applied to medical diagnosis over the past two decades. Traditional statistical methods, particularly Logistic Regression, have been the gold standard due to their interpretability and probabilistic outputs. Logistic Regression models the log-odds of disease presence as a linear combination of features, providing coefficients that indicate the strength and direction of each feature's influence. More recently, ensemble methods such as Random Forests and Gradient Boosting have gained popularity due to their superior predictive performance on complex, non-linear problems. XGBoost, in particular, has achieved state-of-the-art results across numerous medical datasets through its ability to capture intricate feature interactions and handle mixed data types effectively.

2.2. Bayesian Networks in Healthcare

Bayesian Networks have a rich history in medical decision support systems. Early applications include the MUNIN system for diagnosing neuromuscular disorders and the ALARM network for patient monitoring in intensive care units. These systems demonstrated the value of explicitly modelling causal relationships and incorporating expert knowledge.

(Heckerman, 2022) provides a comprehensive tutorial on learning with Bayesian Networks, emphasizing their advantages in handling incomplete data, learning causal relationships, and combining prior knowledge with empirical data. The work demonstrates that Bayesian Networks excel in all common scenarios in medical diagnosis.

Recent work has applied Bayesian Networks to various medical domains including cancer diagnosis, infectious disease modelling, and treatment planning.

2.3. Integration of Large Language Models with Bayesian Networks

Recent research has explored novel integrations between modern AI approaches and traditional probabilistic models. (Feng et al., 2025) introduced BIRD (Bayesian Inference framework for LLMs), which combines Large Language Models with Bayesian Networks to improve probability estimation in decision-making tasks.

3. Methodology

3.1. Datasets

As mentioned in the introduction, this study focuses on two major health care datasets i.e. UCI heart disease dataset and PIMA Indian diabetes dataset. These datasets are well established, used by multiple researchers and have strong performance figures across multiple machine learning models.

3.1.1. UCI Heart Disease dataset

This is a small dataset consisting 303 instances (number of patients), 13 key features and 1 target variable. Among the 13 key features, 5 had continuous values namely age, trestbps (resting blood pressure), chol (serum cholestrol), thalach (maximum heart rate), and oldpeak (depression induced by excercise). The rest of the discrete features are sex, cp (chest pain type), fbs (fasting blood sugar), restecg (resting electrocardiographic results), exang (excercise induced angina), slope (slope of peak exercise), ca (number of major vessels colored by fluoroscopy), and thal (thalassemia). The target variable 'num' indicates the disease severity (0-4) with 0 as no disease and 1-4 as increasing severity. Some instances had a few missing values; those instances were removed from the dataset.

3.1.2. PIMA Indian Diabetes dataset

This dataset consists of 768 patients (records), 8 features and 1 target variable. The features are as follows: pregnancies (number of times pregnant), glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree (genetic predisposition score), and age. The target variable is binary so it indicates as 0 (no diabetes) and 1 (diabetes diagnosed). Some of the features had zero (0) values as which were considered as missing and were filled with the median value of that feature.

3.2. Data pre-processing

The data pre-processing step consists of three steps; first, handling missing and out-of-order values, second, discretization of continuous features and third, splitting the data for training and testing purposes. The explanation of each step is as follows:

3.2.1. Handling missing values

Heart Disease Dataset: Samples with missing values (indicated by '?') were removed, reducing the dataset from 303 to 297 samples. This approach was chosen due to the small proportion of missing data (approximately 2%).

Diabetes Dataset: Zero values in continuous features were replaced with the median value of each feature. Median imputation was selected over mean imputation for robustness to outliers.

3.2.2. Feature Discretization for Bayesian Networks

Bayesian Networks traditionally operate on discrete variables. Continuous features were discretized using the K-Bins Discretizer with the following specifications:

- Number of bins: 3 for heart disease dataset and 2 for diabetes dataset (providing sufficient granularity while avoiding data sparsity)
- Strategy: Quantile-based binning (ensures balanced bin populations)
- Encoding: Ordinal (preserves natural ordering of bins)

3.2.3. Dataset splitting

Both datasets were split into training (80%) and testing (20%) sets using stratified sampling to preserve class distributions. Stratification ensures both sets maintain the original class balance as currently, the ratio of the absence to the presence of disease is not balanced in both datasets.

3.3. Bayesian Network Implementation

3.3.1. Structure Learning

The first step to implement Bayesian network is structure learning. The relationship between the features are established either manually using already established and known domain knowledge or using the **Hill Climb Search** algorithm. The working and performance of subsequent steps depend on how well the structure is formed. A well structure is formed with in-depth understanding of the dataset and problem.

Currently, our approach follows the combination of both manual and Hill Climb search algorithm with the **Bayesian Information Criterion (BIC)** scoring method to learn the structure. The process begins with an empty graph (no edges) and then iteratively optimize the structure by evaluating all possible single-edge modifications (add, delete, reverse) followed by selecting the modification that maximally improves BIC score. This is repeated until no improvement is possible (local optimum). It also ensures the resulting graph is acyclic (DAG property).

BIC Score Formula:

$$BIC = logP(D|G) - (d/2)logN$$

Where:

- P(D|G): Likelihood of data D given graph structure G
- d: Number of parameters in the model
- N: Sample size

The BIC score balances model fit (likelihood) against model complexity (parameter count), preventing over-fitting.

3.3.2. Domain Knowledge Integration

After the automated structure learning, domain knowledge was integrated by adding edges from the already established relationship knowledge in that medical domain. Moreover, domain edges were added only if they did not create cycles and were not contradicted by learned structure (i.e., opposite direction edge not present). For example, for the heart disease dataset:

- cp $\rightarrow$ target (chest pain type strongly predicts heart disease)
- thalach $\rightarrow$ target (maximum heart rate indicates cardiac function)
- exang $\rightarrow$ target (exercise-induced angina is a cardinal symptom)

and for diabetes:

- Glucose $\rightarrow$ Outcome (blood sugar is primary diagnostic criterion)
- BMI $\rightarrow$ Outcome (obesity is major risk factor)
- Age $\rightarrow$ Outcome (risk increases with age)

3.3.3. Parameter Learning

The next step is training the model. We've trained our **DiscreteBayesianNetwork** with the training dataset (features and target combined) using the **Bayesian Estimation** with

BDeu (Bayesian Dirichlet equivalent uniform) prior algorithm.

The model is trained by taking each node X with parents $Pa(X)$ and estimating the conditional probability distribution $P(X|Pa(X))$ [1] using:

$$P(X = x|Pa(X) = pa) = (N_{xpa} + \alpha)/(N_{pa} + \alpha|X|)$$

Where:

- N_{xpa}: Count of samples where X=x and Pa(X)=pa
- N_{pa}: Count of samples where Pa(X)=pa
- α: Prior pseudocount (equivalent sample size / table size)
- $|X|$: Number of states of variable X

3.3.4. Inference

Variable Elimination method is used to get the inference from the learned model and make predictions on the testing dataset.

$$P(\text{Disease}|\text{Observed Features})$$

Process:

- For each test sample, construct evidence dictionary with observed feature values
- Use Variable Elimination to compute posterior probability distribution over target variable
- Select class with maximum posterior probability as prediction
- Record probability of disease class for ROC analysis

3.4. Baseline Methods

Among all the well-established discriminative machine learning models, we've implemented only **logistic regression** and **XGBoost** to compare our model with the baseline performances.

3.4.1. Logistic Regression

The process begins by handling missing values and scaling features using **StandardScalar** followed by training with 1000 max iterations, making predictions on testing dataset and then evaluating the performance.

[1]These formulas are already implemented by the library used and are just mentioned for learning purposes

3.4.2. XGBoost

XGBoost is implemented using XGBClassifier with **logloss** evaluation metric, 0.1 learning rate and 100 n_estimators. Similar to logistic regression, it is also followed by the testing and evaluation steps.

3.5. Evaluation

3.5.1. Evaluation Metrics

The primary metrics used to evaluate the performance of our model's performance are **accuracy** and **AUC-ROC** scores. For better representation and visualization, we've also displayed classification report and confusion matrix. Furthermore, the clinical metrics used are:

- Sensitivity (Recall, TPR):
  True Positive Rate = TP / (TP + FN)
  (Percentage of diseased patients correctly identified)
- Specificity (TNR):
  True Negative Rate = TN / (TN + FP)
  (Percentage of healthy patients correctly identified)
- Precision (PPV):
  Positive Predictive Value = TP / (TP + FP)
  (When model predicts disease, probability of being correct)
- F1-Score:
  Harmonic mean of precision and recall = 2 × (Precision × Recall) / (Precision + Recall)

3.5.2. Cross-Validation

5-fold Stratified Cross-Validation was used to evaluate the performance for each fold. First the we divide data into 5 folds maintaining class proportions. For each fold, we train on 4 folds, validate on 1 fold and record performance metrics. Finally, mean and standard deviation across folds will be reported.

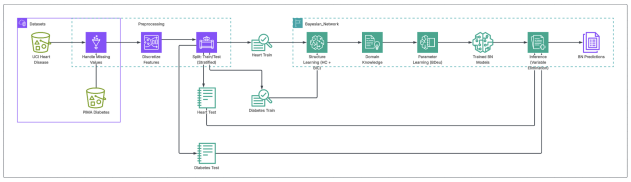

Figure 1. System architecture of our methodology

4. Results

4.1. Heart Disease Dataset

4.1.1. Multiclass Classification (5 Classes)

Class Distribution:

- Class 0 (No disease): 164 samples (55.2%)
- Class 1 (Mild): 55 samples (18.5%)
- Class 2 (Moderate): 36 samples (12.1%)
- Class 3 (Severe): 35 samples (11.8%)
- Class 4 (Very severe): 7 samples (2.4%)

Performance Results:

Multiclass classification proved challenging due to severe class imbalance, particularly for classes 3 and 4. All methods struggled to accurately predict minority classes, with most predictions concentrated on classes 0 and 1. The marginal performance differences between methods suggest that the problem difficulty stems from insufficient data for rare classes rather than model limitations.

Table 1. Performance results for baseline models and Bayesian Network on Heart Disease data set.

Model	Test Acc	CV Acc	AUC-ROC
LR	0.60	0.59 ± 0.03	0.86
XGBoost	0.56	0.56 ± 0.04	0.83
BN	0.61	0.55 ± 0.05	

4.1.2. Binary Classification (Disease vs. No Disease)

After converting classes 1-4 to a single ”disease present” class:

Class Distribution:

- No disease (0): 164 samples (55.2
- Disease present (1): 133 samples (44.8

Performance Results:

Key Findings:

- Binary classification achieves around 25% higher accuracy than multiclass.
- Bayesian Network shows slight edge in accuracy and AUC.

Table 2. Performance results for baseline models and Bayesian Network on Heart Disease data set.

Model	Test Acc	CV Acc	AUC-ROC
LR	0.83	0.82 ± 0.07	0.94
XGBoost	0.83	0.80 ± 0.08	0.91
BN	0.86	0.80 ± 0.06	0.94

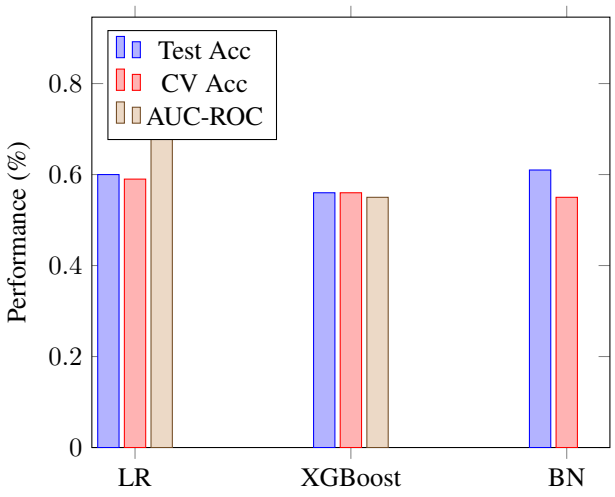

Figure 2. Multi classification performance results on heart disease dataset

- High specificity across all methods (¿83%) indicates low false positive rates.
- Sensitivity range (83-86%) shows good disease detection capability.

4.2. Diabetes Dataset

The PIMA Indian diabetes dataset is a binary classified dataset so only binary classification is performed on this dataset.

Class Distribution:

- No diabetes (0): 500 samples (65.1
- Diabetes (1): 268 samples (34.9

Performance Results:

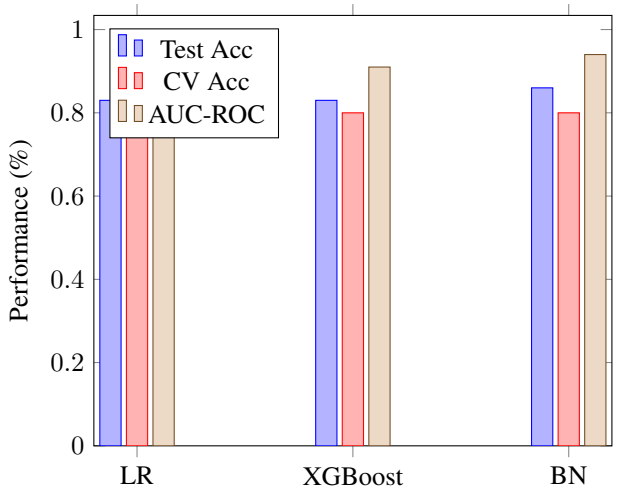

Figure 3. Binary classification performance results on heart disease dataset

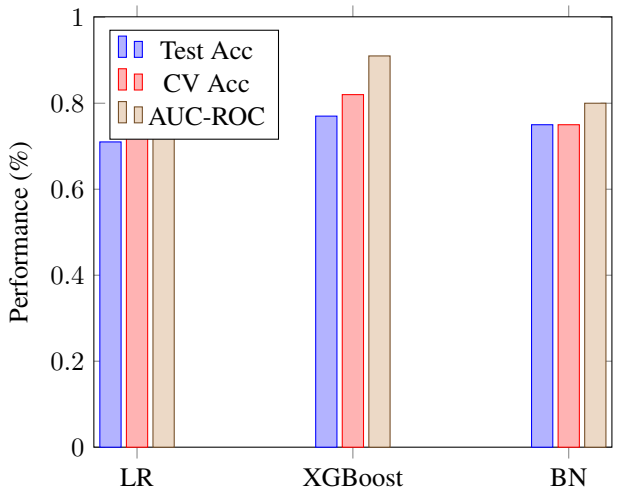

Figure 4. Binary classification performance results on Diabetes dataset

Table 3. Performance results for baseline models and Bayesian Network on Diabetes data set.

MODEL	TEST ACC	CV ACC	AUC-ROC
LR	0.71	0.77 ± 0.01	0.82
XGBOOST	0.77	0.82 ± 0.04	0.91
BN	0.75	0.75 ± 0.03	0.80

Key Findings:

- All methods achieve 71-75% accuracy, consistent with published literature
- XGBoost achieves highest performance across all metrics
- Bayesian Network within 2% of best accuracy
- Specificity consistently higher than sensitivity across methods (easier to identify healthy patients)
- AUC-ROC scores (0.80-0.91) indicate good ranking/risk stratification ability

5. Discussion

5.1. Bayesian Networks: Advantages and Limitations

5.1.1. KEY ADVANTAGES

- Interpretability and Explainability
- Probabilistic Reasoning and Uncertainty Quantification
- Domain Knowledge Integration
- Causal Reasoning
- Bidirectional Inference

5.1.2. KEY LIMITATIONS

- Computational Complexity
- Discretization Information Loss
- Sample Efficiency
- Structure Learning Sensitivity

5.2. Limitations of This Study

- Dataset Size: Both datasets are relatively small (297 and 768 samples), limiting generalization. Larger datasets might reveal greater performance differences.
- Single Dataset per Condition: Results from one heart disease and one diabetes dataset may not generalize to other populations or measurement protocols.
- Discretization Strategy: We used 4-bin quantile discretization; other strategies (domain-informed thresholds, adaptive binning) might improve Bayesian Network performance.
- Structure Learning Algorithm: Hill Climb Search is one of many structure learning approaches; constraint-based methods (PC algorithm) or Bayesian approaches (MCMC) might yield different structures.
- Hyperparameter Tuning Extent: While we tuned key parameters, exhaustive optimization might narrow performance gaps.

- Lack of External Validation: Evaluation used test sets from the same distributions as training data; external validation on different hospital systems would strengthen clinical validity claims.

6. Conclusion

6.1. Summary of Findings

This research implemented and evaluated Bayesian Networks for medical diagnosis across heart disease and diabetes classification tasks, comparing performance against Logistic Regression and XGBoost baselines. Key findings include:

- Competitive Performance: Bayesian Networks achieved 86% (heart disease) and 75.0% (diabetes) accuracy, within 2-3% of best-performing baselines for diabetes dataset, demonstrating that probabilistic graphical models can match discriminative approaches in predictive capability.
- Superior Interpretability: Learned Bayesian Network structures accurately reflected established medical knowledge, with glucose, BMI, chest pain, and cardiac test results emerging as primary predictors consistent with clinical understanding.
- Binary vs. Multiclass: Binary classification substantially outperformed multiclass (83-86% vs. 56-61% accuracy), revealing that disease presence/absence formulation is both more practical clinically and more tractable computationally than fine-grained severity classification with imbalanced classes.
- Domain Knowledge Value: Hybrid structure learning (automated discovery + expert-guided edges) produced clinically valid networks that both fit data and satisfied medical plausibility constraints.
- Computational Trade-off: Bayesian Networks required 20-40x more training time than baselines due to structure learning, though inference time remained acceptable.

6.2. Future Research Directions

This study can be carried out further using different approaches or enhancing the current one to improve the performance. Few of the areas of investigations are:

- Gaussian Bayesian Networks: Eliminate discretization by using continuous probability distributions.
- Dynamic Bayesian Network
- Larger-Scale Studies: Evaluate performance on larger datasets and multiple institutions to assess generalization and scalability.
- Hybrid Models: Combine Bayesian Networks with deep learning (e.g., using neural networks for feature extraction feeding into Bayesian Network structures) to leverage strengths of both models.

6.3. Final Remarks

Bayesian Networks represent a mature yet underutilized approach in medical machine learning. While discriminative models dominate performance benchmarks, the interpretability, uncertainty quantification, and knowledge integration capabilities of Bayesian Networks address critical needs in healthcare AI.

Our results demonstrate that with careful implementation—including appropriate discretization, hybrid structure learning, and Bayesian parameter estimation—Bayesian Networks achieve clinically acceptable accuracy while providing transparency that black-box models cannot match. As healthcare AI moves from research to deployment, the interpretability advantages of Bayesian Networks may prove more valuable than marginal accuracy gains.

The future of medical AI likely lies not in choosing between accuracy and interpretability, but in developing hybrid approaches that achieve both. Bayesian Networks, as one component of a diverse methodological toolkit, offer unique strengths for problems where understanding why a prediction was made is as important as the prediction itself.

References

Feng, Y., Zhou, B., Lin, W., and Roth, D. Bird: A trustworthy bayesian inference framework for large language models, 2025. URL https://arxiv.org/abs/2404.12494.

Heckerman, D. A tutorial on learning with bayesian networks, 2022. URL https://arxiv.org/abs/2002.00269.

Nafar, A., Venable, K. B., Cui, Z., and Kordjamshidi, P. Extracting probabilistic knowledge from large language models for bayesian network parameterization, 2025. URL https://arxiv.org/abs/2505.15918.

A. Important URLS

Github: Medical-Diagnosis-Risk-Scoring-using-Bayesian-Networks

Colab Notebooks:

Heart Binary Bayesian Network

Diabetes Binary Bayesian Network

Heart Multi Bayesian Network

Heart Baseline (Both binary and multi class)

Diabetes Baseline (Binary)

B. Use of AI tools

AI tools (LLMs) were used to enhance productivity in the following ways:

- helping with debugging issues in the codebase by suggesting improvements,
- assisting in LaTeX formatting,
- providing explanations for background Machine learning topics when needed.
- helping with the analysis and creation of charts and models for data